Evaluation: Single Value Metrics (P@n, R-precision and MAP)

COMP3009J: Information Retrieval

Dr. David Lillis (david.lillis@ucd.ie)

UCD School of Computer Science Beijing Dublin International College

Single Value Metrics

- Using Precision and Recall requires two metrics to be calculated.
- We would prefer to have a single metric that captures the overall performance of the system:
 - Precision at n (P@n)
 - R-precision
 - Mean Average Precision (MAP)
 - Binary Preference (bpref)
 - Normalised Discounted Cumulated Gain (NDCG)

Average over several queries

- □ In our discussion of these metrics, we will see how to calculate a score for one set of results for one query.
- Of course, when evaluating an IR system, we should use several queries for evaluation.
- □ In this situation, we must first calculate the score for each query, and then get the average over all the queries.

Precision at n

- \square Sometimes we are interested in the precision amongst the top n results.
- ☐ This is particularly suitable for web search systems where people typically look no further than the first few results.
- □ For example, we may be interested in the precision after 10 documents have been retrieved.
 - Known as "Precision at 10", or "P@10"
- □ This is a measure of the quality of the results that a user is likely to look at.

Precision at n

- □ P@3 = 0.67
 - (i.e. 2 of the top 3 results were relevant)
- □ P @ 10 = 0.4
 - (i.e. 4 of the top 10 results were relevant)

Rank

5

8

9

10

11

12

13

14

15

Document

 d_{123} (r)

 d_{84}

 d_{56} (r)

 d_6

 d_8

d₉ (r)

 d_{511}

 d_{129}

 d_{187}

 d_{25} (r)

 d_{38}

 d_{48}

 d_{250}

 d_{113}

 d_3 (r)

Precision @ n: Problems

- One problem with using P@n is that its performance is affected by the number of relevant documents available.
- For a query that has only 10 relevant documents, a good P@10 score is difficult.
- □ If there are 1,000 relevant documents, a good P@10 score is easy.

R-precision

- A similar metric is **R-precision**, where we are interested in the precision after the top R documents are returned (where R is the number of relevant documents).
- This is similar to P@n except that n is not fixed for all queries (because the number of relevant documents will be different for each query).

R-precision

- R-Precision in our example is the same as P@10, because there are 10 relevant documents for this query.
- ☐ Therefore R-Precision = 0.4

Rank Document

d₁₂₃ (r)

5

8

9

10

11

12

13

14

15

 d_{84}

 d_{56} (r)

 d_6

 d_8

d₉ (r)

 d_{511}

 d_{129}

 d_{187}

 d_{25} (r)

 d_{38}

 d_{48}

 d_{250}

 d_{113}

 d_3 (r)

R-precision

- Another example, with a different set of relevant documents.
- \blacksquare R = {d₃, d₅₆, d₁₂₉}
- R-Precision in this example is the same as P@3, because there are 3 relevant documents for this query.
- Of the first 3 results, 1 is relevant.
- □ Therefore R-Precision = 0.33

Rank Document

d₁₂₃

5

8

9

10

11

12

13

14

15

 d_{84}

 d_{56} (r)

 d_6

 d_8

d₉

 d_{511}

 d_{129} (r)

 d_{187}

 d_{25}

 d_{38}

 d_{48}

 d_{250}

 d_{113}

 d_3 (r)

Mean Average Precision (MAP)

- Mean Average Precision (MAP) has for many years been the most commonly used metric in IR literature to evaluate the performance of systems.
- □ It is a single-value metric based on precision.
- Unlike simple precision, it rewards systems that rank relevant documents at the beginning of the results returned.
- □ Unlike P@10, it continues to examine the later stages of the ranked list, although with lesser weight.

Mean Average Precision (MAP)

- It involves three steps:
 - Firstly, we must calculate the precision at each recall point (at each rank where a relevant document is found).
 - 2. The **Average Precision** for this query is found by dividing the sum of these precision calculations by the total number of relevant documents.
 - 3. For multiple queries, the same procedure must be performed for each. We must calculate the mean of the queries' average precision values, giving us **Mean Average Precision.**

Example

- As before, we assume that there were 10 relevant documents available for retrieval.
- Average Precision (for this query): $\frac{1.0+0.67+0.5+0.4+0.33}{10} = 0.29$
- □ This is the AP for one query.
- To get the Mean Average Precision (MAP), we calculate the average over all queries.

Rank	Document	
1	d ₁₂₃	$P = \frac{1}{1} = 1.00$
2	d ₈₄	
3	d ₅₆	$P = \frac{2}{3} = 0.67$
4	d ₆	
5	d_8	
6	d ₉	$P = \frac{3}{6} = 0.5$
7	d ₅₁₁	
8	d ₁₂₉	
9	d ₁₈₇	
10	d ₂₅	$P = \frac{4}{10} = 0.4$
11	d ₃₈	
12	d ₄₈	
13	d ₂₅₀	
14	d ₁₁₃	
15	d_3	$P = \frac{5}{15} = 0.33$