Evaluation: bpref (Binary Preference)

COMP3009J: Information Retrieval

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

UCD School of Computer Science
Beijing Dublin International College

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. https://doi.org/http://doi.acm.org/10.1145/1008992.1009000

bpref: Binary Preference

- All of the evaluation techniques we have mentioned so far are based on the **Cranfield Paradigm**.
- In this, test **collections** and **queries** are created that have a known set of **relevant documents** associated with them.
- The point is that for each query, **every** document in the collection is judged to be "relevant" or "non-relevant".
- Where this occurs, we say that we have **"complete relevance judgments"**.

bpref: Binary Preference

- With smaller collections, this Cranfield Paradigm is perfect (i.e. complete relevance judgments exist).
- However, as document collections have become larger, complete judgments have become less common and we have **incomplete judgments**. This means that some documents have not been judged so they may or may not be relevant to test queries.
- With large-scale IR collections (such as those based on the web), this complete judgment is impossible to achieve, with potentially millions of documents to be judged for relevance against hundreds of queries.

bpref: Binary Preference

- For the evaluation metrics we have seen so far, they are simplified by **assuming** that unjudged documents are non-relevant.
- It was noticed that many unjudged documents had the effect of lowering evaluation score.
  - **NOTE:** This does not mean that the retrieval was worse: whether a document is relevant or not is not affected by whether somebody has judged it.
- This is a problem because evaluation scores no longer accurately reflect the effectiveness of retrieval.
- For example, what would the effect on the Average Precision score be in our example of nobody had judged document d_9?

If d_9 is judged relevant:

AP = 0.29

If d_9 is unjudged:

AP = 2.24/9 = 0.25

Rank	Document	d_9 judged	d_9 unjudged
1	d_{123}	$P = {}^{1}/_{1} = 1.00$	$P = {}^{1}/_{1} = 1.00$
2	d_{84}		
3	d_{56}	$P = {}^{2}/_{3} = 0.67$	$P = {}^{2}/_{3} = 0.67$
4	d_6		
5	d_8		
6	d_9	$P = {}^{3}/_{6} = 0.5$	
7	d_{511}		
8	d_{129}		
9	d_{187}		
10	d_{25}	$P = {}^{4}/_{10} = 0.4$	$P = {}^{3}/_{10} = 0.3$
11	d_{38}		
12	d_{48}		
13	d_{250}		
14	d_{113}		
15	d_3	$P = {}^{5}/_{15} = 0.33$	$P = {}^{4}/_{15} = 0.27$

bpref: Binary Preference

- The idea behind **bpref** is that these unjudged documents should not impact so largely on the evaluation score.
  - From the paper: Buckley & Voorhees, "Retrieval Evaluation with Incomplete Information", SIGIR 2004
- bpref ignores documents that have not been judged.
- The only documents considered are those that were **judged relevant** or **judged non-relevant**.
- This is an accepted metric for large-scale IR systems where complete relevance judgments are impossible to achieve.

bpref: Binary Preference

- bpref, for a query with *R* relevant documents is calculated as:
  - $B = \frac{1}{R}\sum_{r \in R} 1 - \frac{|n \; ranked \; higher \; than \; r|}{R}$
    - where *n* is a member of the first *R* judged non-relevant documents.

- In other words:
  - For each relevant document in that was retrieved in the answer set:
    - Count the number of non-relevant documents above it in the result set (this is $|n \; ranked \; higher \; than \; r|$). This cannot be greater than the total number of relevant documents (R)
    - The score for that document is $1 - \frac{|n \; ranked \; higher \; than \; r|}{R}$
  - Average this score over all the judged relevant documents.

bpref Calculation

- Key:
  - R: Relevant
  - N: Non-Relevant
  - U: Unjudged
- Assume a total of 4 available judged relevant documents (R=4).
- bpref = (0.75+0.75+0+0)/4 = 0.375

Document	bPref Contribution
N	
R	$1 - {}^{1}/_{4} = 0.75$
U	
R	$1 - {}^{1}/_{4} = 0.75$
U	
N	
N	
N	
R	$1 - {}^{4}/_{4} = 0$
N	
R	$1 - {}^{4}/_{4} = 0$

bpref: Binary Preference

- This works well in most situations.
- However, when *R* is very small (i.e. there are only one or two relevant documents) it fails.
- To overcome this problem, we can instead use bpref-10, which is given by:
  - $bpref10 = \frac{1}{R}\sum_{r \in R} 1 - \frac{|n\ ranked\ higher\ than\ r|}{10+R}$
    - where *n* is a member of the first 10+*R* judged non-relevant documents.

Effect of bpref

- From Buckley and Voorhees 2004 it can be seen that when complete judgments are available, there is no noticeable difference between MAP and bpref-10.
- As the judgments become less complete, bpref is more stable than the others.