MSCC-IRWS
Information Retrieval and Web Search
Evaluation 1

Jason Farina

Faculty of Computing Science
Griffith College Dublin

jason.farina@gcd.ie

Introduction

- Evaluation is concerned with "How well does the system work?"
- There are many measurable quantities for this:
  1. Processing: How quickly does the user receive a response? How well are resources utilised?
  2. User Experience: Does the user enjoy using the system?
  3. Search: How effective is the system in satisfying the information need?
- Question 3 is of most interest in this lecture. This evaluates the actual retrieval algorithms that we are using.

Introduction

- Evaluation of the effectiveness of an IR system (particularly in the research area) is a vital topic
- There are many different techniques used in Information Retrieval and there need to be accepted means to quantify their performance
- Many metrics exist to do this.
- We will look at the following commonly used metrics:
  1. Precision / Recall
  2. Interpolated Precision
  3. Mean Average Precision (MAP)
  4. bPref

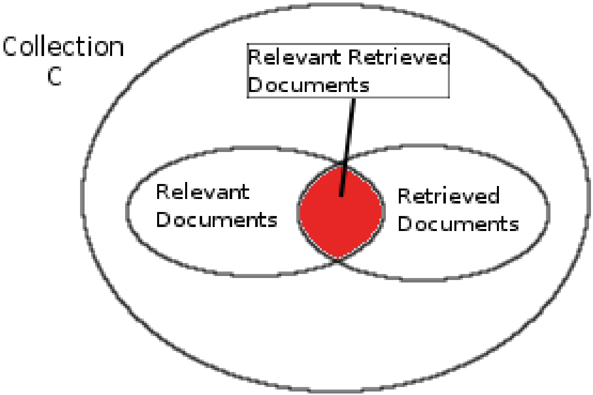
Introduction: The Cranfield Paradigm

- Evaluation in IR typically follows the *Cranfield Paradigm*.
- This requires three parts:
  - A standard document collection (e.g. the Cranfield collection of 1400 abstracts from aeronautical engineering articles).
  - A set of standard queries, each representing some information need.
  - A set of *relevance judgments* for every query.

Introduction: The Cranfield Paradigm

- A *relevant* document is one that (at least partially) satisfies a user's information need.
- Unfortunately, an information need is a very subjective thing.
- As an alternative, we use experts to judge whether each document is relevant to each query (the Cranfield experiments used aeronautical engineers).
- The next slide shows how the answer set (returned by an IR system), the relevant set (the set of documents judged to be relevant) and the collection are combined.

Relevance



RelRet:	Relevant Retrieved
Rel:	Relevant
Ret:	Retrieved

Introduction

- The answer set is not really a set.
- Generally, it is in the form of a ranked list
  - The Boolean Model is an exception to this!
- The purpose of evaluating the effectiveness of an IR technique is to evaluate the quality of this ranked list.

Example

from Modern Information Retrieval

- Consider a query, q, on document collection C where $|C| = 800$
- $Rel = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$
- The ranked list of retrieved documents, RET, is given by:

1. d_{123} (r)
2. d_{84}
3. d_{56} (r)
4. d_6
5. d_8
6. d_9 (r)
7. d_{511}
8. d_{129}
9. d_{187}
10. d_{25} (r)
11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_3 (r)

Precision / Recall

- Precision and Recall are the two most basic (and most widely used) evaluation metrics in IR, upon which many others are based.
- **Precision** is the fraction of the set of retrieved documents (RET) that are relevant (i.e. they are also in REL)
  - $Precision = \frac{|RelRet|}{|RET|}$
- **Recall** is the fraction of the relevant documents (REL) that have been retrieved (i.e. they are also in RET)
  - $Recall = \frac{|RelRet|}{|REL|}$

Example

- Looking at the example we can see:
  - the number of relevant documents for the query is $|REL| = 10$
  - the number of retrieved documents for the query is $|RET| = 15$
  - the number of relevant documents in the retrieved set is $|RELRET| = 5$
- Hence the precision is: $\frac{RELRET}{RET} = \frac{5}{15} = 33\%$
- The recall is: $\frac{RELRET}{REL} = \frac{5}{10} = 50\%$

Precision vs Recall

- The two metrics of precision and recall are often inversely related: as one increases the other decreases.
- Precision will be high whenever a system is good at avoiding non-relevant documents.
  - A system can achieve very high precision by retrieving very few documents.
- Recall will be high whenever a system finds many relevant documents.
  - A system can achieve 100% recall by retrieving all the documents in the collection.

Precision vs Recall

- Which one is the most important?
- That is task dependent!
- As an example: if a user searches the web they want *high precision* i.e. they want as many of the returned documents to be relevant.
- There are very many documents that will help satisfy the information need, so the user does not need to examine all of them (i.e. recall does not need to be high).
- Instead, the user wishes to avoid trawling through non-relevant documents.

Precision vs Recall

- On the other hand, a patent lawyer researching a patent must ensure that they get all of the relevant documents and hence they want *high recall*.
- If any relevant documents are missed, this may have serious consequences so it is essential that all relevant documents are returned.
- They will tolerate lower precision to facilitate this (i.e. they are more likely to be willing to read through some non-relevant documents).
- Ideally every IR system would have both recall and precision of 100% (the answer set is equal to the relevant set).

Interpolated Precision

Precision / Recall Graphs

- Precision and Recall are set-based measures
- In a ranked list we can measure the precision at each *recall point*.
  - We begin at the top of the ranked results list.
  - Recall increases when a relevant document is retrieved: this is a *recall point*.
  - We calculate precision at each recall point, using only the documents returned so far as our answer set.
    - So $|RET|$ is the number of documents retrieved *so far*.
    - ... and $|RELRET|$ is the number of relevant documents retrieved *so far*.

Example

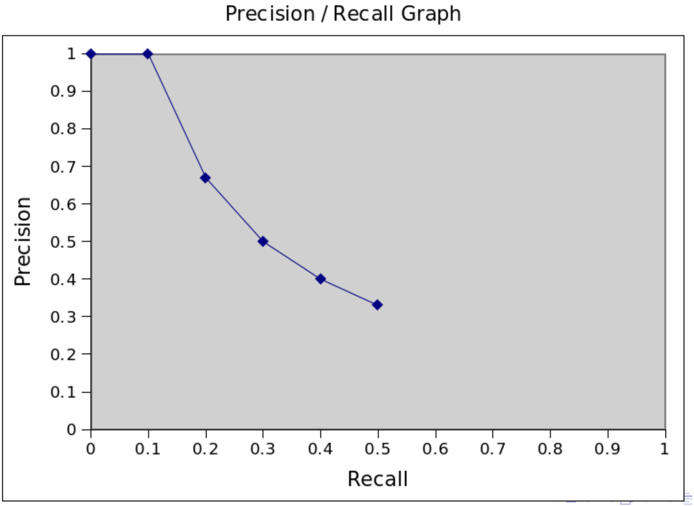
(a reminder)

- $Rel = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$

1. d_{123} (r)
2. d_{84}
3. d_{56} (r)
4. d_6
5. d_8
6. d_9 (r)
7. d_{511}
8. d_{129}
9. d_{187}
10. d_{25} (r)
11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_3 (r)

- In this example, we would need to calculate precision 5 times: at position 1, 3, 6, 10 and 15.
- e.g. at position 6:
  - $Precision = \frac{3}{6} = 0.5$
  - $Recall = \frac{3}{10} = 0.3$

Precision / Recall graph



Interpolated Precision

- For the previous example, there were 10 relevant documents, so each time we found a new relevant document, we had found an extra 10% of all the relevant document.
- Strictly speaking we should calculate the precision / recall at every retrieved document in the ranked list.
- This can lead to a messy 'shark's fin' type graph
- The interpolated precision graph smooths this out, showing interpolated precision at 11 standard levels of recall: 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100%.

Intepolated Precision

- Interpolation gives us a standard graph to compare against other sets of results, by giving us precision measurements for each of the standard recall point, rather than the actual recall levels that appear in the result set.
- The interpolated precision at each standard recall point is the maximum precision that is found between that point and the next one.
- For example, in calculating the interpolated precision at a recall level of 10%, we would use the maximum precision found at any recall point between 10% and 20%. If there are no known recall points in that range, we would use the precision at 20% recall (which may itself be interpolated).

Interpolated Precision Procedure

- Calculate precision at all recall points
- Working from 100% to 0%, for each standard recall level
    - If there is no recall point at this recall level or higher, precision is 0%
    - If there is a recall point at this recall level or higher, that is the precision at this recall point.
    - If there are multiple recall points at this recall level or higher, take the highest precision.

Example, revisited

- How has the interpolated precision/recall graph changed by changing the relevant documents?
- $Rel = \{d_3, d_{56}, d_{129}, d_4\}$
- The ranked list of retrieved documents, RET, is given by:

1. d_{123}
2. d_{84}
3. d_{56} (r)
4. d_6
5. d_8
6. d_9
7. d_{511}
8. d_{129} (r)
9. d_{187}
10. d_{25}
11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_3 (r)

Interpolated Precision/Recall

1. d_{123}
2. d_{84}
3. d_{56} (r)
4. d_{6}
5. d_{8}
6. d_{9}
7. d_{511}
8. d_{129} (r)
9. d_{187}
10. d_{25}
11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_{3} (r)

- We have three recall points: at positions 3, 8 and 15.
- At position 3:
  - $Precision = \frac{1}{3} = 0.33$
  - $Recall = \frac{1}{4} = 0.25$

Interpolated Precision/Recall

1. d_{123}
2. d_{84}
3. d_{56} (r)
4. d_{6}
5. d_{8}
6. d_{9}
7. d_{511}
8. d_{129} (r)
9. d_{187}
10. d_{25}
11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_{3} (r)

- At position 8:
  - $Precision = \frac{2}{8} = 0.25$
  - $Recall = \frac{2}{4} = 0.5$
- At position 15:
  - $Precision = \frac{3}{15} = 0.2$
  - $Recall = \frac{3}{4} = 0.75$

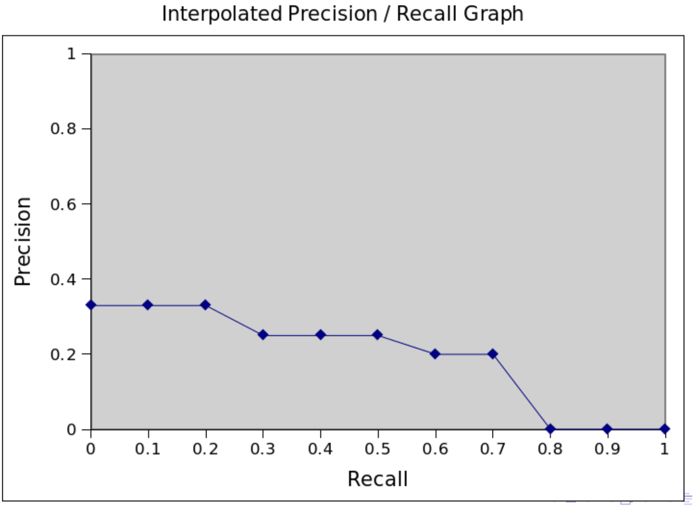
Interpolated Precision/Recall

- So we have precision calculated at three recall levels:
    - P=0.33 @ R=0.25
    - P=0.25 @ R=0.5
    - P=0.2 @ R=0.75
- We begin interpolating for a standard recall level of 100%. As the system never returned the fourth relevant document, precision here is 0.
- For recall level 90% and 80% it is also 0 (no recall point was found beyond those recall levels)
- For recall levels 70% and 60%, there is only one recall point higher, so precision there is 0.2.
- For recall level 50% there are now two recall points to take into account, so precision is the larger of these (i.e. 0.25).

Interpolated Precision/Recall

- Reminder:
  - P=0.33 @ R=0.25
  - P=0.25 @ R=0.5
  - P=0.2 @ R=0.75
- This is also the case for recall levels 40% and 30%.
- On reaching a recall level of 20%, a new recall point becomes relevant. Again we take the highest precision value, which in this case is 0.33.
- This is also the appropriate precision value for recall levels 10% and 0%.

Interpolated Precision / Recall graph



Precision at n / R-Precision

- Sometimes we are interested in the precision amongst the top n results
- This is particularly suited to web search systems as people look no further than the first few results
- For instance we may be interested in the precision after 3 documents have been retrieved
- In the case of our example system this value is 0.66 as two of the first three documents were relevant to the query
- Similarly we might be interested in the precision at a particular recall level
- For instance the R-precision at $R = 0.3$ in our original example is 0.5.

Summary

- Done:
  - Basic precision / recall
  - Interpolated precision
  - Precision at *n* documents
  - R-precision
- To Do:
  - Mean Average Precision
  - bPref
  - Which metric to choose??