Information Retrieval

Computer Science Tripos Part II

Evaluation

Stephen Clark
Natural Language and Information Processing (NLIP) Group
Introduction to Evaluation

• We want to know how well a retrieval system performs

• What is “performance” in an IR setting?
  – For a DBMS, performance is data retrieval time, since search is exact
  – For an IR system, search is inexact
    * still interested in retrieval time
    * also interested in retrieval accuracy
    * may be interested in other factors: ease of use, financial, presentation of documents, help in formulating queries, ...

• IR evaluation has focused primarily on retrieval accuracy: how good is a system at returning documents which are relevant to the user need?
History

- Evaluation has been a key issue in IR since the 60’s
  - consequence of the empirical approach taken to IR
- Early work compared manual vs. automatic indexing
- The TREC competitions (over the last decade) have been very influential
Difficulties with IR Evaluation

- “Relevance” is difficult to define precisely
  - who makes the judgement?
  - humans are not very consistent

- Information need may not be clear – so how can we determine if it’s been satisfied?

- Difficult to separate the user from the system, especially in interactive retrieval

- Judgements depend on more than just document and query

- For large document collections, difficult to determine the set of relevant documents
Evaluation under Laboratory Conditions

- Evaluation has been used as an analytical tool in an experimental setting
  - e.g. to determine if one weighting scheme is better than another
  - implies control of experimental variables

- Abstraction of IR system from operational setup

- Largely ignored interaction with the user

- Concentration on measures like precision and recall using standard test collections
TREC

- Text Retrieval Conference
  - Established in 1992; annual conference
  - designed to evaluate large-scale IR
    (2 gigabyte document collections, up to a million documents)
  - Run by NIST (US technology agency)
  - In 1992 25 organisations – industrial and academic – participated
  - In 2003 93 groups participated from 22 different countries
  - http://trec.nist.gov/
Format of TREC

• TREC consists of IR research tracks
  – ad-hoc, filtering, cross-language, genomics, HARD, interactive, question-answering, terabyte, video, web
  * HARD: High Accuracy Retrieval from Documents; uses information about, and interaction with, user

• Timetable:
  – Spring: researchers train/develop systems
  – Summer: system is run on final test collection and results submitted to NIST for evaluation
  – November: conference takes place to compare results

• Competition encourages research and enables successful approaches to be adopted for the next round
  – does it work?
Test Collections

- Test collections used to compare retrieval performance of systems / techniques
  - set of documents
  - set of queries (or topics)
    * typically text description of user need, or information request, from which final query is constructed
  - set of relevance judgements

- How to compare performance?
  - results (set of returned documents, usually ranked) compared using some performance measure
    - precision and recall most common measures

- Ideally use multiple test collections
  - performance can be collection-specific
Use of Test Collections

• Before TREC, IR testing was on a relatively small scale

• Earlier work tended to use the same test material to maintain comparability

• Large test collections (both queries and documents) are important
  – to ensure statistical significance of results
  – to convince commercial system operators of the validity of the results

• TREC tracks typically have hundreds of thousands of documents, and hundreds of topics
**Sample TREC Query**

<num> Number: 508
<desc> hair loss is a symptom of what diseases

<desc> Description:
Find diseases for which hair loss is a symptom.

<narr> Narrative:
A document is relevant if it positively connects the loss of head hair in humans with a specific disease. In this context, “thinning hair” and “hair loss” are synonymous. Loss of body and/or facial hair is irrelevant, as is hair loss caused by drug therapy.
Humans decide which document–query pairs are relevant.
Determining Relevant Documents

• Did the system return all possible relevant documents?
  – need a relevance judgement for every document in the collection, for every query/topic
  – at 30s a document/topic pair, would take 6,500 hours to judge 800,000 TREC documents for one topic

• TREC solution is pooling
  – select $N$ runs per system
  – take the top $K$ (usually 100) documents returned by each system (according to system’s ranking) for those runs
  – then assume all relevant documents are in union and manually assess this set
  – pooling found not to be bias towards systems contributing to the pool
**Precision and Recall for Document Retrieval**

- **Precision** = $|Ra|/|A|$  
  - precision = $\hat{P}(\text{relevant} | \text{retrieved})$

- **Recall** = $|Ra|/|R|$  
  - recall = $\hat{P}(\text{retrieved} | \text{relevant})$
Another Representation

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>not retrieved</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

- **precision** = \( \frac{A}{A+B} \)
  - \( \hat{P}(relevant|retrieved) \)

- **recall** = \( \frac{A}{A+C} \)
  - \( \hat{P}(retrieved|relevant) \)

- **miss** = \( \frac{C}{A+C} \)
  - \( \hat{P}(not-retrieved|relevant) \)

- **false alarm** (fallout) = \( \frac{B}{B+D} \)
  - \( \hat{P}(retrieved|not-relevant) \)
Recall-precision curve

- Plotting precision and recall (versus no. of documents retrieved) shows inverse relationship between precision and recall
- Precision/recall cross-over can be used as combined evaluation measure

- Plotting precision versus recall gives recall-precision curve
- Area under normalised recall-precision curve can be used as evaluation measure
Recall-criticality and precision-criticality

- Inverse relationship between precision and recall forces general systems to go for compromise between them.

- But some tasks particularly need good precision whereas others need good recall:

<table>
<thead>
<tr>
<th>Precision-critical task</th>
<th>Recall-critical task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Little time available</td>
<td>Time matters less</td>
</tr>
<tr>
<td>A small set of relevant documents answers the information need</td>
<td>One cannot afford to miss a single document</td>
</tr>
<tr>
<td>Potentially many documents might fill the information need (redundantly)</td>
<td>Need to see each relevant document</td>
</tr>
<tr>
<td>Example: web search for factual information</td>
<td>Example: patent search</td>
</tr>
</tbody>
</table>
Single Value Measures

- F-score \( = \frac{1}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P+R} \)

- F-score is **harmonic mean** of P and R: inverse of average of inverses

- F-score is 1 when \( P = R = 1 \) and 0 when \( P \) or \( R \) are 0

- Penalises low values of P or R
  - it is very easy to obtain high precision (just return very few documents) or high recall (return all documents)
**Geometric Interpretation of F-Measure**

- **P** = \( \frac{|A \cap B|}{|A|} \)
- **R** = \( \frac{|A \cap B|}{|B|} \)

\[
F = \frac{2PR}{(P + R)}
\]

\[
= \frac{2\left(\frac{|A \cap B|^2}{|A| \cdot |B|}\right)}{(|A \cap B| \left(\frac{1}{|A|} + \frac{1}{|B|}\right))}
\]

\[
= \frac{2 |A \cap B|}{|A| + |B|}
\]
Single Value Measures

- **E-measure** = \( \frac{1}{\alpha P^\frac{1}{\beta} + (1-\alpha) R^\frac{1}{\beta}} \)

  - used to emphasis precision or recall
    - weighted harmonic mean of precision and recall
    - high \( \alpha \) emphasises precision

- Transforming by \( \alpha = \frac{1}{\beta^2 + 1} \) gives
  \[ E = \frac{(\beta^2 + 1) PR}{\beta^2 P + R} \]

- \( \beta = 1 \) (\( \alpha = \frac{1}{2} \)) gives F-score

- \( \beta > 1 \) emphasises precision; \( \beta < 1 \) emphasises recall
Measure for Ranked Retrieval

- Precision and Recall well defined for sets
- But matching can be defined as a matter of degree
  - vector space model returns similarity score for each document
- How to evaluate the quality of the rank-ordering, as well as the number and proportion of relevant documents retrieved?
**Precision/Recall @ Rank**

1. $d_{12}$ 2. $d_{123}$ 3. $d_4$ 4. $d_{57}$ 5. $d_{157}$ 6. $d_{222}$ 7. $d_4$ 8. $d_{26}$ 9. $d_{77}$ 10. $d_{90}$

- Suppose there are 3 relevant documents
  - **P@n**: $P@3 = 0.33$, $P@5 = 0.2$, $P@8 = 0.25$
  - **R@n**: $R@3 = 0.33$, $R@5 = 0.33$, $R@8 = 0.66$

- Ranks chosen for reporting depend on expected quantity of documents retrieved

- Rank statistics give some indication of how quickly user will find relevant documents from ranked list

- But may want to abstract away from ranking, since size of ranking will depend on query and document set
**Precision at Recall** $r$

### Ranking 1:
- **Recall:** 0.2, 0.4, 0.4, 0.6, 0.6, 0.8, 1.0
- **Precision:** 1.0, 0.5, 0.67, 0.5, 0.4, 0.38, 0.44, 0.5

### Ranking 2:
- **Recall:** 0.0, 0.2, 0.2, 0.4, 0.6, 0.8, 1.0, 1.0, 1.0
- **Precision:** 0.0, 0.5, 0.33, 0.25, 0.4, 0.57, 0.63, 0.55, 0.5

- **r1:** $p @ r 0.2 = 1.0$; $p @ r 0.4 = 0.67$; $p @ r 0.6 = 0.5$; $p @ r 0.8 = 0.44$; $p @ r 1.0 = 0.5$

- **r2:** $p @ r 0.2 = 0.5$; $p @ r 0.4 = 0.4$; $p @ r 0.6 = 0.5$; $p @ r 0.8 = 0.57$; $p @ r 1.0 = 0.63$
Single Value Summary

- Useful to have a single number effectiveness measure
  - easy to read and interpret
  - may want to optimise for a machine learning algorithm

- **Average precision** is popular in IR

ranking 1:

| recall | 0.2 | 0.2 | 0.4 | 0.4 | 0.4 | 0.6 | 0.6 | 0.6 | 0.8 | 1.0 |
| precision | 1.0 | 0.5 | 0.67 | 0.5 | 0.4 | 0.5 | 0.43 | 0.38 | 0.44 | 0.5 |

av prec = 0.62

ranking 2:

| recall | 0.0 | 0.2 | 0.2 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 | 1.0 | 1.0 |
| precision | 0.0 | 0.5 | 0.33 | 0.25 | 0.4 | 0.5 | 0.57 | 0.63 | 0.55 | 0.5 |

av prec = 0.52
Single Value Summary

- Previous measure was average precision at seen relevant documents

- TREC average precision also accounts for any relevant documents not retrieved

- Suppose there are 8 relevant documents in total (3 are not retrieved by either system)
  - av. prec for r1: \((1 + 0.67 + 0.5 + 0.44 + 0.5)/8 = 0.39\)

- So TREC average precision also has a recall component, in that it considers all relevant documents
Averaging over Queries

- Need an evaluation measure over more than one query

- Average precision over queries for standard recall levels (0.1, 0.2, 0.3, ..., 1.0)?

- But \(|Ra|/|R|\) rarely seen at these levels
  - if only 3 relevant documents, recall can only be 0.33, 0.67.

- Answer: interpolate between actual recall values to get average precision at standard recall levels
  - many possibilities for interpolation; see Modern Information Retrieval, Ch. 3
TREC’s Single Value Summary

- **Average precision** for a single query is the mean of the precision after each relevant document is retrieved

- **Mean average precision** for a set of queries is the mean of the average precision scores for each query
  - popular single value metric to represent system performance over a complete query / document set
IR Performance

- Difficult to raise performance in both precision and recall (precision/recall trade-off)
  - any improvement in precision typically results in a decrease in recall, and vice versa

- Even with small collections, difficult to raise performance beyond 40%/40% P/R level

- With larger collections 30%/30% is more likely

- Systems using statistically based natural language indexing provide respectable performance which is hard to beat
Summary

- Focused on evaluation for ad-hoc retrieval
  - other issues arise when evaluating different tracks, e.g. QA, although typically still use P/R-based measures

- Evaluation for **interactive** tasks is more involved

- Significance testing is an issue
  - could a good result have occurred by chance?
  - is the result robust across different document sets?
  - slowly becoming more common
  - underlying population distributions unknown, so apply weak tests such as the sign test
Readings for Today

- Relevant parts of the course textbook
- Modern Information Retrieval, Ch. 3
- Readings in Information Retrieval, Ch. 4
- Information Retrieval (van Rijsbergen), Ch. 7