Introduction

Framework for the evaluation of an IR system:
- **test collection** consisting of (i) a document collection, (ii) a test suite of information needs and (iii) a set of relevance judgements for each doc-query pair
- **gold-standard judgement of relevance**
  → classification of a document either as relevant or as irrelevant wrt an information need
- **NB**: the test collection must cover at least 50 information needs

Overview

Standard test collections
Evaluation for unranked retrieval
Evaluation for ranked retrieval
Assessing relevance
System quality and user utility
Conclusion

Standard test collections

- **Cranfield collection**: 1398 abstracts of journal articles about aerodynamics, gathered in UK in the 1950s, plus 255 queries and exhaustive relevance judgements
- **TREC (Text REtrieval Conference)**: collection maintained by the US National Institute of Standards and Technology since 1992
  - TREC Ad Hoc Track: test collection used for 8 evaluation campaigns led from 1992 to 1999, contains 1.89 million documents and relevance judgements for 450 topics
  - TREC 6-8: test collection providing 150 information needs over 528000 newswires
  → current state-of-the-art test collection
  → note that the relevance judgements are not exhaustive
Standard test collections (continued)
- **GOV2**: collection also maintained by the NIST, containing 25 millions of webpages (larger than other test collections, but smaller than current collection supported by WWW search engines)
- **NTCIR** (Nii Test Collection for IR systems): various test collections focusing on East Asian languages, mainly used for cross-language IR
- **CLEF** (Cross Language Evaluation Forum): collection focusing on European languages
  [http://www.clef-campaign.org](http://www.clef-campaign.org)
- **REUTERS**: Reuters 21578 and REUTERS RCV1 containing respectively 21,578 newswire articles and 806,791 documents, mainly used for text classification

Evaluation for unranked retrieval: basics
- 2 basic effectiveness measures: precision and recall
  
  \[
  \text{Precision} = \frac{\# \text{relevant retrieved}}{\# \text{retrieved}} \\
  \text{Recall} = \frac{\# \text{relevant retrieved}}{\# \text{relevant}}
  \]
- In other terms:

<table>
<thead>
<tr>
<th>Relevant</th>
<th>Not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>true positive</td>
</tr>
<tr>
<td>Not retrieved</td>
<td>false negative</td>
</tr>
</tbody>
</table>

  \[
  \text{Precision} = \frac{tp}{tp + fp} \\
  \text{Recall} = \frac{tp}{tp + fn}
  \]

Evaluation for unranked retrieval: basics (continued)
- **Accuracy**: proportion of the classification relevant/not relevant that is correct
  
  \[
  \text{accuracy} = \frac{tp + tn}{tp + fp + tn + fn}
  \]

  Problem: 99.9% of the collection is usually not relevant to a given query (potential high rate of false positives)
- Recall and precision are inter-dependent measures:
  - precision usually decreases while the number of retrieved documents increases
  - recall increases while the number of retrieved documents increases
Evaluation for unranked retrieval: F-measure

- Measure relating precision and recall:
  \[ F = \frac{1}{\alpha \times \frac{1}{Pr} + (1 - \alpha) \times \frac{1}{Re}} \]

- When the coefficient \( \alpha = 0.5 \):
  \[ F = \frac{2 \times Pr \times Re}{Pr + Re} \]

- Uses a harmonic mean rather than an arithmetic one for dealing with extreme values

Overview

Standard test collections
Evaluation for unranked retrieval
Evaluation for ranked retrieval
Assessing relevance
System quality and user utility
Conclusion

Evaluation for ranked retrieval

- NB1: precision, recall and F-measure are set-based measures (order of documents not taken into account)

- NB2: if we consider the first \( k \) retrieved documents, we can compute the precision and recall values
  - we can plot the relation between precision and recall for each value of \( k \)
  - if the \((k+1)\)th is not relevant \( \Rightarrow \) recall is the same, but precision decreases
  - if the \((k+1)\)th is relevant \( \Rightarrow \) recall and precision increase

Evaluation for ranked retrieval (continued)

- Precision-recall curve:
Evaluation for ranked retrieval (continued)

- Precision-recall curve:

\[ \text{Precision-recall curve:} \]

- Interpolation of the precision (smoothing):

\[ P_{\text{inter}}(r) = \max_{r' \geq r} P(r') \]

Ranked retrieval: efficiency measures

- **11-point interpolated average precision**: For each information need, the value \( P_{\text{inter}} \) is measured for the 11 recall values 0.0, 0.1, 0.2, \ldots 1.0. The arithmetic mean of \( P_{\text{inter}} \) for a given recall value over the information needs is then computed.

\[ \text{Average 11-point precision/recall graph (from Manning et al, 2008)} \]

Ranked retrieval: efficiency measures (continued)

- **Mean Average Precision (MAP)**:

For an information need, the average precision is the arithmetic mean of the precisions for the set of top \( k \) documents retrieved after each relevant document is retrieved.

\[ q_j \in Q : \text{information need} \]
\[ \{d_1, \ldots, d_l\} : \text{relevant documents for } q_j \]
\[ R_{kj} : \text{set of ranked retrieved document from top to } d_k \]
\[ \text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{kj}) \]

NB: when \( d_i \) (1 ≤ i ≤ j) is not retrieved, \( \text{Precision}(R_{ij}) = 0 \)

Ranked retrieval: efficiency measures (continued)

- **Precision at k**:

For www search engines, we are interested in the proportion of good results among the \( k \) first answers (say the first 3 pages).

- **Pros**: does not need an estimate of the set of relevant documents
- **Cons**: unstable measure, does not average well
**Ranked retrieval: efficiency measures (continued)**

- **R-precision**:  
  Alternative for precision at k. R-precision refers to the best precision on the precision/recall curve when considering a set of relevant documents and looking at the first answers  
  \[ Pr = \frac{r}{Re} \]
  where \( r \) is the number of relevant retrieved documents  
  NB: also called break-even point

- **Normalized Discounted Cumulative Gain (NDCG)**:  
  Evaluation made for the top \( k \) results  
  \[ NDCG(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{Q} Z_k \sum_{m=1}^{k} \frac{2^{R(j,m)} - 1}{\log(1 + m)} \]
  where:  
  \( R(j, d) \) is the score given by assessors to document \( d \) for query \( j \) (NB: non-binary notion of relevance)  
  \( Z_k \) is a normalization factor (perfect ranking at \( k = 1 \))

**Overview**

- Standard test collections
- Evaluation for unranked retrieval
- Evaluation for ranked retrieval
- Assessing relevance
- System quality and user utility
- Conclusion

**Assessing relevance**

- How good is an IR system at satisfying an information need?  
  Needs an agreement between judges  
  - computable via the kappa statistic:  
    \[ \text{kappa} = \frac{P(A) - P(E)}{1 - P(E)} \]
    where:  
    \( P(A) \) is the proportion of agreements within the judgements  
    \( P(E) \) is the proportion of expected agreements  
  - In case of binary decisions (e.g. relevant vs not relevant),  
    \( P(E) = 0.5 \)  
  - Otherwise, a marginal statistic is used (see below)
Assessing relevance: an example
Consider the following judgements (from Manning et al., 2008):

<table>
<thead>
<tr>
<th>Judge 1</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>300</td>
<td>20</td>
<td>320</td>
</tr>
<tr>
<td>No</td>
<td>10</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>90</td>
<td>400</td>
</tr>
</tbody>
</table>

\[
P(A) = \frac{370}{400} \quad P(\text{rel}) = \frac{320 + 310}{800} \quad P(\text{notrel}) = \frac{80 + 90}{800}
\]

\[
P(E) = P(\text{rel})^2 + P(\text{notrel})^2 \quad k = 0.776
\]

Assessing relevance (continued)

- Interpretation of the kappa statistic $k$:
  - $k \geq 0.8$ good agreement
  - $0.67 \leq k < 0.8$ fair agreement
  - $k < 0.67$ bad agreement
- Note that the kappa statistic can be negative if the agreements between judgements are worse than random
- In case of large variations between judgements, one can choose an assessor as a gold-standard
  - considerable impact on the absolute assessment
  - little impact on the relative assessment
Assessing relevance: limits

- Limit of the assessment: does the relevance of a document change in context (i.e. when the document appears after some other document)?
- Case of www search engines: duplicates
- How to evaluate the marginal relevance?
  → evaluating diversity or novelty.

Overview

Standard test collections
Evaluation for unranked retrieval
Evaluation for ranked retrieval
Assessing relevance
System quality and user utility

System quality and user utility

- Ultimate interest: how satisfied is the user with the results the system gives for each of its information needs?
- Evaluation criteria for an IR system:
  - fast indexing
  - fast searching
  - expressivity of the query language
  - size of the collection supported
  - user interface (clearness of the input form and of the output list, e.g. snippets, etc)

System quality and user utility (continued)

- Quantifying user happiness?
  - for www search engines: "do the users find the information they are looking for?" can be quantified by evaluating the proportion of users getting back to the engine
  - for intranet search engines: this efficiency can be evaluated by the time spent searching for a given piece of information
  - general case: user studies evaluating the adequacy of the search engine with the expected usage (eCommerce, etc)
Refining a deployed system

- **A/B test**: Standard test for evaluating changes, particularly used for **www search engines**
  - concerns the modification of 1 functionality between 2 variants A and B of a system
  - 10% of the users are *blindly* redirected to the new variant B
  - the results of the two systems A and B are then studied
- Example: ranking algorithm
  - comparison of the result clicking between users of system A and those of system B

Overview

- Standard test collections
- Evaluation for unranked retrieval
- Evaluation for ranked retrieval
- Assessing relevance
- System quality and user utility

Conclusion

- Standard test collections: Cranfield, TREC, CLEF, etc.
- Evaluation for unranked retrieval: *precision, recall* and *f-measure*
- Evaluation for ranked retrieval: 11-point interpolation, *Mean Average Precision, R-precision* and *Normalized Discounted Cumulative Gain*
- Assessing relevance: *kappa statistic*
- Discussion about system quality, user utility and system refining

References

- C. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval
- Cyril Cleverdon *The significance of the Cranfield tests on index languages* (1991)
  [http://citeseer.ist.psu.edu/buckley00evaluating.html](http://citeseer.ist.psu.edu/buckley00evaluating.html)
- Chris Buckley *the trec_eval software*