Evaluation of Retrieval Systems

Performance Criteria
1. Expressiveness of query language
   • Can query language capture information needs?
2. Quality of search results
   • Relevance to users’ information needs
3. Usability
   • Search Interface
   • Results page format
   • Other?
4. Efficiency
   – Speed affects usability
   – Overall efficiency affects cost of operation
5. Other?

Quantitative evaluation

• Concentrate on quality of search results
• Goals for measure
  – Capture relevance to user information need
  – Allow comparison between results of different systems
• Measures define for sets of documents returned
• More generally “document” could be any information object

Core measures: Precision and Recall

• Need binary evaluation by human judge of each retrieved document as relevant/irrelevant
• Need know complete set of relevant documents within collection being searched
• Recall = \frac{\# relevant documents retrieved}{\# relevant documents}
• Precision = \frac{\# relevant documents retrieved}{\# retrieved documents}

Combine recall and precision

F-score (aka F-measure) defined to be: harmonic mean of precision and recall

\[ \text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

Use in “modern times”

• Defined in 1950s
• For small collections, these make sense
• For large collections,
  – Rarely know complete set relevant documents
  – Rarely could return complete set relevant documents
• For large collections
  – Rank returned documents
  – Use ranking!

\[ \hat{h} \text{ satisfies } (n-h)/n = (h-m)/m. \text{ Also } (1/m) - (1/h) = (1/h) - (1/m) \]
Ranked result list

• At any point along ranked list
  – Can look at precision so far
  – Can look at recall so far
    • if know total # relevant docs
    • Google’s “about N results” inadequate estimate
  • Can focus on points that relevant docs appear
    – If m\textsuperscript{th} doc in ranking is k\textsuperscript{th} relevant doc so far, precision is k/m
    • No a priori ranking on relevant docs

Plot: precision versus recall

• Choose standard recall levels: r\textsubscript{1}, r\textsubscript{2} …
  – Eg 10%, 20% …
  – Define “precision at recall level r\textsubscript{j}”
    \[ p(r\textsubscript{j}) = \text{max over all } r \text{ with } r_{\text{sr}} < r\textsubscript{j} + 1 \text{ of precision} \text{ when recall } r \text{ achieved} \]

Single number characterizations I

• Can look at precision at one fixed critical position of ranking: "\textit{Precision at } k\text"
  – If know are T relevant documents can choose k=T
  • May not want to look that far even if know T
  – Can choose set of R relevant docs, and calc. precision at k=R only with respect to these docs
  • "R-precision" of Intro IR
  • can only do with some prior analysis of collection
  – For Web search
    • Choose k to be number pages people look at
    • k=? What expecting?

Single number characterizations II

1) Record precision at each point a relevant document encountered through ranked list
  • Don’t need know all relevant docs
  • Can cut off ranked list at predetermined rank
  2) Average the recorded precisions in (1)
    \[ = \text{average precision for a query result} \]

\textbf{Mean Average Precision (MAP):}
For a set of test queries, take the mean (i.e. average) Of the average precision for each query
• Compare retrieval systems with MAP

Single number characterizations III

\textbf{Reciprocal rank:}
Capture how early get relevant result in ranking
reciprocal rank of ranked results of a query
\[
\frac{1}{\text{rank of highest ranking relevant result}}
\]
• perfect = 1 \rightarrow worse \rightarrow 0
•  = average precision if only one relevant document
get \textit{mean reciprocal rank} of set of test queries

Beyond binary relevance

• Sense of degree to which document satisfies query
  – classes, e.g: excellent, good, fair, poor, irrelevant
• Can look at measures class by class
  – limit analysis to just excellent docs?
  – combine after evaluate results for each class
• Need new measure to capture all together
  – does \textit{document ranking match} "excellent, good, fair, poor, irrelevant" rating?
Discounted cumulative gain (DCG)

• Assign a gain value to each relevance class
  – e.g. 0 (irrel.), 1, 2, 3, 4 (best) assessor’s score
  – how much difference between values?
    – text uses \( (2^{\text{assessor’s score}} - 1) \)

• Let \( d_1, d_2, \ldots, d_k \) be returned docs in rank order
  • \( G(i) = \text{gain value of } d_i \)
    – determined by relevance class of \( d_i \)

• \( \text{DCG}(i) = \sum_{j=1}^{i} \left( \frac{G(j)}{\log_b (1+j)} \right) \)
  – parameter \( b \): how much doc retrieved lower down in ranking is penalized – text uses \( b=2 \)

Using Discounted Cumulative Gain can compare retrieval systems on query by

• plotting values of \( \text{DCG}(i) \) versus \( i \) for each
  – plot gives sense of progress along rank list

• choosing fixed \( k \) and comparing \( \text{DCG}(k) \)
  – if one system returns \( < k \) docs, fill in at bottom with “irrel”

• can average over multiple queries
  – text “Normalized Discounted Cumulative Gain”
    • normalized so best score for a query is 1

Comparing orderings

Two retrieval systems both return \( k \) excellent documents. How different are rankings?

• Measure for two orderings of \( n \)-item list: 
  Kendall’s Tau

  inversion: pair of items ordered differently in the two orderings

  Kendall’s Tau (order1, order2) = 
  \[ 1 - \left( \frac{\text{# inversions}}{\frac{1}{2} n(n-1)} \right) \]

Using Measures

• Statistical significance versus meaningfulness
  • Use more than one measure

• Need some set of relevant docs even if don’t have complete set

  How?
  – Look at TREC studies