Evaluation of Information Retrieval Systems

Anoop Kunchukuttan
Roll No. 06305407
Under the Guidance Of
Prof. Soumen Chakrabarti

Department of Computer Science and Engineering,
Indian Institute of Technology, Bombay
Mumbai

4 Dec 2006

Retrieval Evaluation

- **What is Relevance?**
  - Difficult to define
  - Subjective, Personal
  - Diverse user needs
- **Need to quantify relevance**
  - For comparing IR systems
  - For having objective criteria for system selection

Outline of Seminar

- Performance measures for IR systems
- Text Retrieval Conference
- XML Retrieval Evaluation and INEX
- Implications of Retrieval to Ranking

User Needs

- Resource Finding
- Specific Answer
- Broad Topic Search
- Browsing

IR evaluation must correctly identify the user need.

Basic Tasks

- Quantification of relevance
- Building a test collection
- Ensure completeness of relevance judgements

Precision and Recall

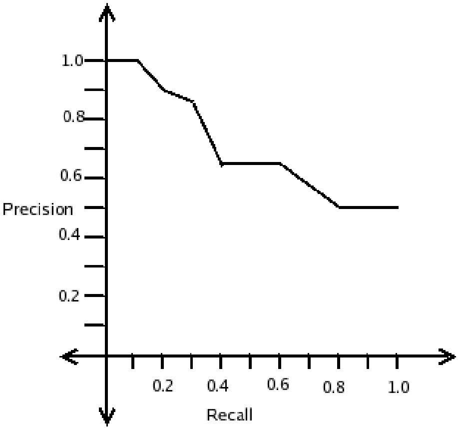
- Fraction of relevant documents retrieved

$$\text{Recall} = \frac{TP}{TP + FN}$$

- Fraction of the documents retrieved that are relevant

$$\text{Precision} = \frac{TP}{TP + FP}$$

- *Standard Recall Values* 0%, 10%, 20%, ... 100% relevant docs retrieved.
- *Interpolated Precision* $P(r_j) = \max_{r_i \le r \le r_{j+1}} P(r)$



Justification for Precision-Recall

- **The Probabilty Ranking Principle (PRP)**
  'Order documents in decreasing order of probability of relevance to user'
- The system tries to maximize

$$\text{logit}\ \phi(d_i) = \log \frac{\theta_1(d_i)}{\theta_2(d_i)} + \text{logit}\ \gamma$$

where,
logit $p = \frac{\log p}{\log(1-p)}$
$\theta_1 =$ P(doc retrieved|doc relevant)
$\theta_2 =$ P(doc retrieved|doc non-relevant)
$\phi =$ P(doc relevant|doc retrieved)
$\gamma =$ P(document relevant)

θ_1 is recall, θ_2 is fallout and ϕ is precision

PR based Metrics

- **Precision@N**: Precision after N documents retrieved
- **R-Precision**: Precision after all R relevant documents retrieved
- **F-1 score**: Harmonic mean of precision and recall

$$F1 = \frac{2PR}{(P + R)}$$

- **Mean Average Precision (MAP)**: Average of the precision values at each relevant document retrieved, across different queries.
- **PRBEP**: Precision-Recall Break Even Point

Pros and Cons

Pros

- Easy to compute with linear ordering among documents
- Justified as a metric for PRP based ranking

Cons

- Does not address different kinds of user needs
- Batch mode metric
- Relevance judgement costly for large corpus
- Addresses binary relevance only

Normalized Discounted Cumulated Gain

- Supports multiple levels of relevance
- Each relevance level assigned a grade level
- **Cumulated Gain**

$$CG[i] = \begin{cases} G[1] & \text{if } i = 0 \\ CG[i-1] + G[i] & \text{otherwise} \end{cases}$$

Example

$$G = [2, 3, 3, 2, 2, 3, 3, 1]$$

Cumulated gain vector

$$CG = [2, 5, 8, 10, 12, 15, 18, 19]$$

Normalized Discounted Cumulated Gain(2)

- Give more importance to elements at higher ranks.
- Use a discount factor to achieve this
- **Discounted Cumulated Gain**

$$DCG[i] = \begin{cases} CG[i] & \text{if } i < b \\ DCG[i-1] + \frac{G[i]}{\log_b i} & \text{if } i \geq b \end{cases}$$

- **NDGC**: Normalize DCG with respect to the ideal DCG score

$$NDCG[i] = \frac{DCG[i]}{DCG_{ideal}[i]}$$

Some other Metrics

- **Mean reciprocal rank**: Reciprocal of the rank of the highest ranked relevant document.
- **bpref**: Based on relative ranking rather than absolute ranking, it is a function of number of times non-relevant docs are retrieved before relevant documents.

$$bpref = \frac{1}{R} \sum_{r} 1 - \frac{|\text{n ranked higher than r}|}{R}$$

- **%no-metric**: Fraction of queries on which the system returned no relevant results in the top-10

Text Retrieval Conference

- Annual conference sponsored by NIST and the US DoD since 1992
- To encourage research in information retrieval based on large test collections
- Development of evaluation methodologies
- Undertakes a number of focussed track like Web, HARD, Robust Retrieval, Terabyte, etc. for specific tasks.

TREC Methodology

- Define retrieval tasks and topics (queries)
- Provide test collection
- Submit results to NIST
- Create reference results (relevance judgement)
- Benchmark systems against standard results using defined evaluation measures (eg. MAP, NDGC).

Pooling in TREC

- Relevance Judgement method used by TREC
- Avoids exhaustive assessment
- Pool N (say 100) top results for query from each retrieval system
- Assess the results in this reduced pool
- Documents not in pool considered not relevant
- Works well for small and medium test collections

Some TREC Tracks

HARD Track

- Focussed on high precision retrieval
- Systems exploit relevance feedback from user
- Ternary relevance scheme
- Effectiveness metric: R-Precision

Robust Retrieval Track

- Improving effectiveness of poorly performing queries
- Emphasizes a system's least effective topics
- Ternary relevance scheme
- Effectiveness metric: Gmap (Geometric Mean Average Precision)
- Gmap more sensitive to low scores

XML Relevance

Different from flat text relevance due to explicit document structure

- Fine grained information
- Two dimensional view of relevance
  - Exhaustivity
  - Specificity
- Graded relevance
- Consistency of results

Precision/Recall for XML?

- Don't support multiple levels of relevance
- Can't measure exhaustivity/specificity
- Don't consider overlap of a component with other result elements

INEX

- Initiative for Evaluation of XML Retrieval
- Annual conference, started in 2002
- Aim: To develop approaches to XML retrieval evaluation
- Mainly focussed on content-oriented XML
- Evaluation methodology very similar to TREC

Relevance Grades

- Relevance grades along two dimensions
- Exhaustivity grades (0-3): Not exhaustive, Marginally exhaustive, Fairly exhaustive, Highly exhaustive
- Specificity grades (0-3): Not specific, Marginally specific, Fairly specific, Highly specific
- Quantise into a single score:

$$quant_{gen}(e, s) = e.s$$

where,
e=exhaustivity of element
s=specificity of element

Extended Cumulated Gain Metrics

Computing Gain Values
The simple case (Overlaps not considered)

$$xG[i] = rv[c_i] = quant(assess(c_i))$$

where,
$rv[c_i]$ is the relevance value of element c_i
assess is a function which returns the (e, s) values of element c_i
Considering component overlaps

$$rv(c_i) = \begin{cases} quant(assess(c_i)) & c_i \text{ not seen} \\ (1-\alpha).quant(assess(c_i)) & c_i \text{seen completely} \\ \alpha.\frac{\sum_{j=1}^{m}(rv(c_j)).|c_j|}{c_i} + (1-\alpha).quant(assess(c_i)) & c_i \text{ seen partially} \end{cases}$$

Extended Cumulated Gain Metrics (2)

- **The xCG metric:**

$$xCG[i] = \sum_{j=1}^{i-1} xG[j]$$

- **The normalized nxCG metric:**

$$nxCG[i] = \frac{xCG[i]}{xCG_{ideal}[i]}$$

Gain-Recall/Effort-Precision

- **Gain-Recall Value:**
  Cumulated gain value divided by the total achievable cumulated gain. It is analogous to recall.

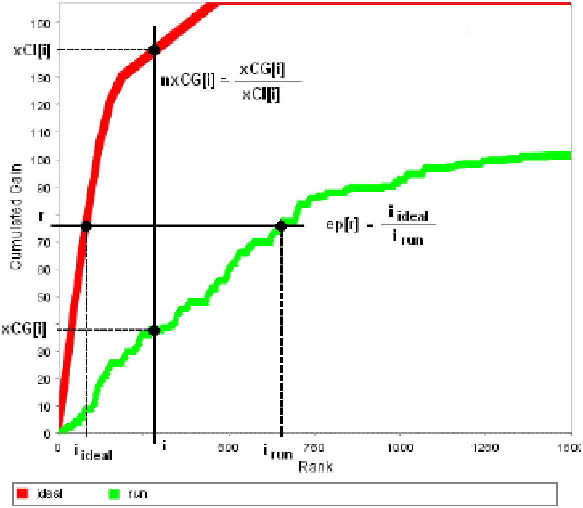
$$gr[i] = \frac{xCG[i]}{xCG_{ideal}[n]}$$

- **Effort-Precision:**
  Estimate how much effort the user has to undergo to reach a particular gain-recall level relative to the ideal gain vector. It is analogous to precision.

$$ep[r] = \frac{iideal}{irun}$$

where,
$iideal$ = ideal curve's rank position at which the cumulated gain is r
$irun$ = the rank position at which the cumulated gain of r is reached by evaluated system

Gain-Recall/Effort-Precision Graph



Comparable to traditional PR graphs

Exploiting Evaluation for Ranking

- Retrieval Evaluation not part of design
- Relevance Feedback important source of user information
- Can relevance information be used to improve rankings?
  - Optimizing for an evaluation metric
  - Using relevance feedback to optimize ranking

Expected Metric Principle

- Make the evaluation metric the quantity to be optimized
- The metric should reflect the user need
- *'In a probabilistic context, one should directly optimize for the expected value of the metric of interest'*

An Example-The 1-call metric

Maximize the chances of getting atleast one relevant result in the top n results

$$Pr[r_0 \cup r_1 \cup ... r_{n-1} | d_0, d_1, ..., d_{n_1}].$$

For the case of n=2

$$\begin{aligned}
& Pr[r_0 \cup r_1 | d_0, d_1] \\
& = Pr[r_0 | d_0, d_1] + Pr[r_1 \cap \neg r_0 | d_0, d_1] \\
& = Pr[r_0 | d_0, d_1] + Pr[r_1 | d_0, d_1, \neg r_0] Pr[\neg r_0 | d_0, d_1] \\
& = Pr[r_0 | d_0] + Pr[r_1 | d_0, d_1, \neg r_0] Pr[\neg r_0 | d_0]
\end{aligned}$$

This suggests heuristic for the k-call case:

1. Select the first document based on its relevance, as in PRP.
2. Now select the most relevant document considering only the rest of the documents and assuming that the already retrieved documents are not relevant.

Side-effect: Promotes diversity in top-k ranks

Optimizing ranking using relevance feedback

- Relevance feedback important source of user behaviour
- Clickthrough, time spent on a result page, scrolling of result page, etc. etc
- Cheap to collect this information
- Clickthrough most indicative feedback.

Clickthrough as relevance feedback

- Judgement based on top-k results, summary of results
- Only relative judgement
- **Presentation bias**: User may not always click the links due to relevance alone.
- Remove background noise in the clickthrough data to get bias free distribution.

$$o(q, r, f) = C(f) + rel(q, r, f)$$

where,
o is observed value of a user feature f for query q and result r
$C(f)$ is background component
rel is the relevance component

Using clickthrough to rank

- Pairwise relevance information can be extracted
- Interpreting clickthrough
  - Skip Above: Results above clicked result less relevant
  - Skip Next: Clicked result more relevant than next result
- Optimizing rankings
  - Re-rank top k results
  - Use relevance information as feature in base ranker

Issues in using relevance feedback

- Does the user interaction element provide relevance feedback and can it be quantified?
- Does it need to be pre-processed?
- How will relevance information be extracted?

Concluding Remarks

- One size doesn't fit all
- Retrieval measures address user's sensibilities
- Incomplete relevance judgements a challenge
- Exploit relevance feedback to optimize results
- Approaches to design systems to optimize for the right metrics

THANK YOU

Some References

- TREC publications. http://trec.nist.gov/pubs.html.
- Agichtein, Brill, Dumais, . Improving web search ranking by incorporating user behavior information. *ACM SIGIR conference on Research and development in information retrieval*, 2006.
- Agichtein, Brill, Dumais, Ragno. Learning user interaction models for predicting web search result preferences. *ACM SIGIR conference on Research and development in information retrieval*, 2006.
- Baeza-Yates and Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Longman Publishing Co. Inc, 1st edition, 1999.
- Kalervo Jarvelin and Jaana Kekalainen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4), 2002.

Some References

- Thorsten Joachims. Optimizing search engines using clickthrough data. Technical report, Cornell University, Department Of Computer Science, 2002.
- David Karger and Harr Chen. Less is More: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of the ACM SIGIR*, 2006.
- Kazai and Lalma. INEX 2005 evaluation metrics, 2005. http://inex.is.informatik.uni-duisburg.de/2005/inex-2005-metricsv6.pdf.
- Benjamin Piwowarski and Mounia Lalmas. Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. *In CIKM 04:Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 2004.