Introduction to Information Retrieval

Evaluation
Rank-Based Measures

- Binary relevance
  - Precision@K (P@K)
  - Mean Average Precision (MAP)
  - Mean Reciprocal Rank (MRR)

- Multiple levels of relevance
  - Normalized Discounted Cumulative Gain (NDCG)
Precision@K

- Set a rank threshold K
- Compute % relevant in top K
- Ignores documents ranked lower than K

Ex: 
- Prec@3 of 2/3
- Prec@4 of 2/4
- Prec@5 of 3/5
Mean Average Precision

- Consider rank position of each relevant doc
  - $K_1, K_2, \ldots K_R$

- Compute Precision@$K$ for each $K_1, K_2, \ldots K_R$

- Average precision = average of P@$K$

- Ex: \[
\begin{bmatrix}
\text{Green} & \text{Red} & \text{Green} & \text{Green} \end{bmatrix}
\]
  has AvgPrec of \[
\frac{1}{3} \cdot \left( \frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76
\]

- MAP is Average Precision across multiple queries/rankings
Average Precision

Ranking #1: \( \frac{1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6}{6} = 0.78 \)

Ranking #2: \( \frac{0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6}{6} = 0.52 \)
MAP

average precision query 1 \( = \frac{(1.0 + 0.67 + 0.5 + 0.44 + 0.5)}{5} = 0.62 \)

average precision query 2 \( = \frac{(0.5 + 0.4 + 0.43)}{3} = 0.44 \)

mean average precision \( = \frac{(0.62 + 0.44)}{2} = 0.53 \)
Mean average precision

- If a relevant document never gets retrieved, we assume the precision corresponding to that relevant doc to be zero.
- MAP is macro-averaging: each query counts equally.
- Now perhaps most commonly used measure in research papers.
- Good for web search?
- MAP assumes user is interested in finding many relevant documents for each query.
- MAP requires many relevance judgments in text collection.
When There’s only 1 Relevant Document

- Scenarios:
  - known-item search
  - navigational queries
  - looking for a fact

- Search Length = Rank of the answer
  - measures a user’s effort
Mean Reciprocal Rank

- Consider rank position, $K$, of first relevant doc

- Reciprocal Rank score $= \frac{1}{K}$

- MRR is the mean RR across multiple queries
Critique of pure relevance

- **Relevance vs Marginal Relevance**
  - A document can be redundant even if it is highly relevant
    - Duplicates
    - The same information from different sources
  - Marginal relevance is a better measure of utility for the user
    - But harder to create evaluation set
    - See Carbonell and Goldstein (1998)

- Using facts/entities as evaluation unit can more directly measure true recall

- Also related is seeking diversity in first page results
  - See *Diversity in Document Retrieval* workshops
Introduction to Information Retrieval

Yahoo! Search

Also try: toyota safety ratings, toyota safety recall, More...

Toyota
- Recall
  Toyota Takes Care of its Customers. Read the FAQs at Toyota.com.
  www.Toyota.com/Recall

- Safety
  www.Toyota.Edmunds.com

TOYOTA | Car Safety Innovation and Technology
Toyota home page for car safety and car technology Prius model.
www.safetytoyota.com - Cached

- home page for car safety and car technology...
  We are presenting Toyota's safety technologies for cars. We clearly explain about
car safety and car technology using movies and more.
  www.safetytoyota.com/en_gb - Cached

Toyota Safety Ratings - Toyota Safety Features - Motor Trend ...
Motor Trend offers Toyota safety ratings, comprehensive auto safety reports, and more.
View all of the standard Toyota safety features ...
motorwear.com/new_cars/Toyota/safety_ratings/index.html - 149k - Cached

Toyota Motor Europe Corporate Site Safety
Our approach Toyota believes that all stakeholders in the road safety equation share a
responsibility to reduce the frequency of road accidents ...
www.toyota.eu/Safety - Cached

European Safety Brochure 2005
4047k - Adobe PDF - View as html
not guarantee that all accidents or injuries will be avoided when driving a Toyota and/or
Lexus brand motor vehicle equipped with the safety systems ...
www.toyota.no/Images/Safety_Brochure_tcm003-344451.pdf

Toyota - Star Safety System
Star Safety System ... Toyota Mobility Program. Careers. Contact Us. Home. contact us.
site map your privacy rights. legal terms. Toyota Newsroom. sign up for info ...
www.toyota.com/vehicles/democrats/star-safety.html - 56k - Cached

Toyota Prius Safety Ratings - CarsDirect
Get overall safety ratings and NHTSA crash test results for the Toyota Prius at
CarsDirect.
Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks

- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant document
  - the lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined
Discounted Cumulative Gain

- Uses *graded relevance* as a measure of usefulness, or *gain*, from examining a document.
- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks.
- Typical discount is $1/\log (rank)$.
  - With base 2, the discount at rank 4 is $1/2$, and at rank 8 it is $1/3$. 
Summarize a Ranking: DCG

- What if relevance judgments are in a scale of [0, r]? r > 2
- Cumulative Gain (CG) at rank n
  - Let the ratings of the n documents be \( r_1, r_2, \ldots r_n \) (in ranked order)
  - \( CG = r_1 + r_2 + \ldots r_n \)
- Discounted Cumulative Gain (DCG) at rank n
  - \( DCG = r_1 + \frac{r_2}{\log_2 2} + \frac{r_3}{\log_2 3} + \ldots \frac{r_n}{\log_2 n} \)
  - We may use any base for the logarithm, e.g., base=b
Discounted Cumulative Gain

- $DCG$ is the total gain accumulated at a particular rank $p$:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- Alternative formulation:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log(1+i)}$$

- used by some web search companies
- emphasis on retrieving highly relevant documents
DCG Example

• 10 ranked documents judged on 0-3 relevance scale:
  3, 2, 3, 0, 0, 1, 2, 2, 3, 0

• discounted gain:
  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0
  \[= 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0\]

• DCG:
  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61
Summarize a Ranking: NDCG

- Normalized Cumulative Gain (NDCG) at rank n
  - Normalize DCG at rank n by the DCG value at rank n of the ideal ranking
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc
  - Compute the precision (at rank) where each (new) relevant document is retrieved => \( p(1), \ldots, p(k) \), if we have k rel. docs

- NDCG is now quite popular in evaluating Web search
NDCG - Example

4 documents: \(d_1, d_2, d_3, d_4\)

<table>
<thead>
<tr>
<th>i</th>
<th>Ground Truth</th>
<th>Ranking Function(_1)</th>
<th>Ranking Function(_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Document Order</td>
<td>(r_i)</td>
<td>Document Order</td>
</tr>
<tr>
<td>1</td>
<td>d(_4)</td>
<td>2</td>
<td>d(_3)</td>
</tr>
<tr>
<td>2</td>
<td>d(_3)</td>
<td>2</td>
<td>d(_4)</td>
</tr>
<tr>
<td>3</td>
<td>d(_2)</td>
<td>1</td>
<td>d(_2)</td>
</tr>
<tr>
<td>4</td>
<td>d(_1)</td>
<td>0</td>
<td>d(_1)</td>
</tr>
</tbody>
</table>

\(\text{NDCG}_{\text{GT}} = 1.00\) \(\text{NDCG}_{\text{RF}_1} = 1.00\) \(\text{NDCG}_{\text{RF}_2} = 0.9203\)

\[
\text{DCG}_{\text{GT}} = 2 + \left( \frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309
\]

\[
\text{DCG}_{\text{RF}_1} = 2 + \left( \frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309
\]

\[
\text{DCG}_{\text{RF}_2} = 2 + \left( \frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619
\]

\[\text{MaxDCG} = \text{DCG}_{\text{GT}} = 4.6309\]
Introduction to Information Retrieval

**Out of 4728 rel docs, we’ve got 3212**

**Recall=3212/4728**

**Precision@10docs**

about 5.5 docs in the top 10 docs are relevant

**Breakeven Point (prec=recall)**

**Mean Avg. Precision (MAP)**
What Query Averaging Hides

Slide from Doug Oard’s presentation, originally from Ellen Voorhees’ presentation