Introduction to

Information Retrieval

Chris Manning, Pandu Nayak and Prabhakar Raghavan

CS276: Information Retrieval and Web Search

Evaluation of IR systems

edited by Danilo Croce

WmIR course AA2015/2016

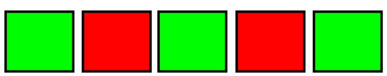
University of Roma Tor Vergata

Rank-Based Measures

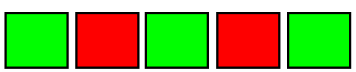
- Binary relevance
  - Precision@K (P@K)
  - Mean Average Precision (MAP)
  - Mean Reciprocal Rank (MRR)

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

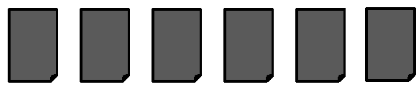
Precision@K

- 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

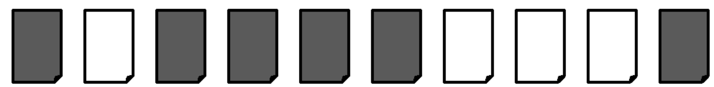
Mean Average Precision

- 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:  has AvgPrec of $$\frac{1}{3}\times\left(\frac{1}{1}+\frac{2}{3}+\frac{3}{5}\right)\approx 0.76$$
- MAP is Average Precision across multiple queries/ rankings

Average Precision

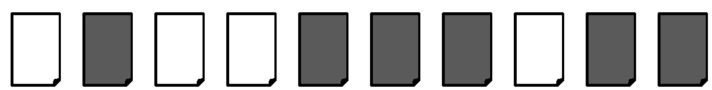
 = the relevant documents

Ranking #1



Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2

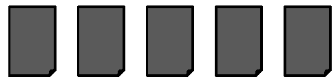


Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

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

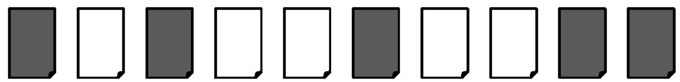
$$\text{Ranking \#2: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

MAP

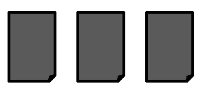


= relevant documents for query 1

Ranking #1

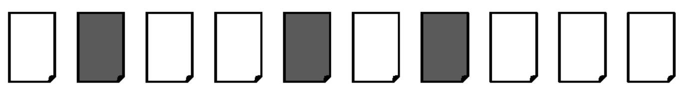


Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5



= relevant documents for query 2

Ranking #2



Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

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

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

$$\textit{mean average precision} = (0.62 + 0.44)/2 = 0.53$$

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

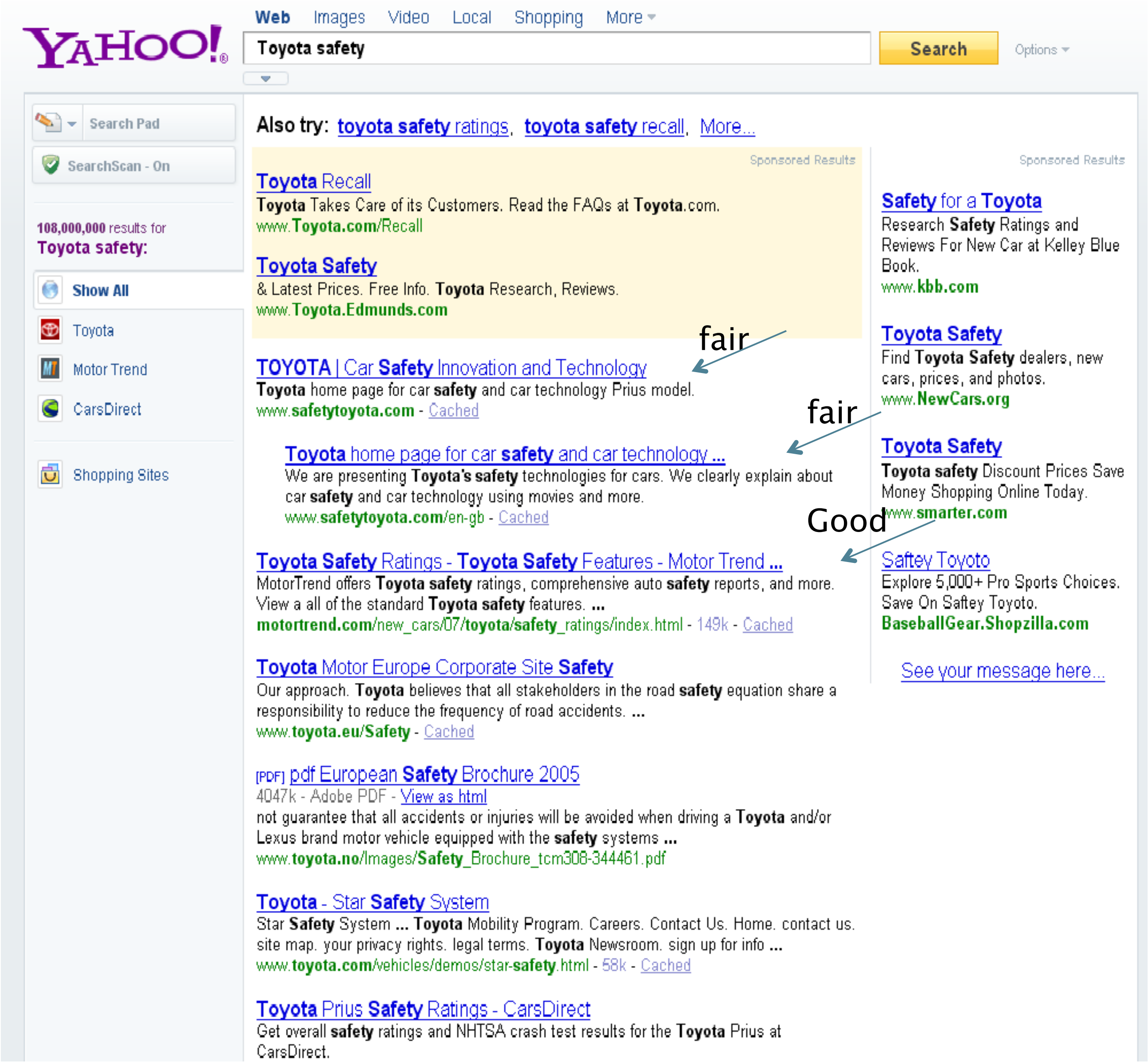
- Scenarios:
  - known-item search
  - navigational queries
  - looking for a fact
- Search Length = Rank of the answer
  - measures a user's effort

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



fair

fair

Good

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, …r_n (in ranked order)
  - CG = $r_1+r_2+\ldots r_n$
- Discounted Cumulative Gain (DCG) at rank n
  - DCG = $r_1 + r_2/\log_2 2 + r_3/\log_2 3 + \ldots\ r_n/\log_2 n$
    - 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

Discounted Cumulative Gain Example

- 10 ranked documents judged on 0-3 relevance scale:
  - 3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- discounted gain (DG):
  - 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), …,p(k), if we have k rel. docs
- NDCG is now quite popular in evaluating Web search

Summarize a Ranking: NDCG

- Perfect ranking:
  - 3, 3, 3, 2, 2, 2, 1, 0, 0, 0
- ideal DCG values:
  - 3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10
- Actual DCG:
  - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61
- NDCG values (divide actual by ideal):
  - 1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88
- NDCG $\leq$ 1 at any rank position