Advances in IR Evaluation

Ben Carterette    Evangelos Kanoulas    Emine Yilmaz
Course Outline

• Intro to evaluation
  – Evaluation methods, test collections, measures, comparable evaluation

• Low cost evaluation

• Advanced user models
  – Web search models, novelty & diversity, sessions

• Reliability
  – Significance tests, reusability

• Other evaluation setups
Low-Cost Evaluation (4)

• Estimating *measures* with less judgments
  – Aslam et al. SIGIR06, Yilmaz and Aslam CIKM06, Yilmaz et al SIGIR09

• Estimating systems *ranking* with less judgments
  – Carterette et al. SIGIR06, Moffat et al. SIGIR07
Goals for a Test Collection

• Different goals suggest different approaches:
  – Find the relevant documents:
    • Pooling
    • Move-to-Front pooling, Hedge
    • Interactive Searching and Judging
  – Estimate the value of an evaluation measure:
    • infAP, xinfAP, statAP
  – Compare two or more systems by some measure:
    • MTC (Minimal Test Collections)
MTC (Minimal Test Collections)

• MTC is an adaptive, episodic algorithm for deciding which documents to judge

• Its goals:
  – Accurately compare two or more systems
  – Make a minimum number of judgments
  – Use existing judgments to help choose

• Not goals of MTC:
  – Select documents most likely to be relevant
  – Find all (or even most) of the relevant documents
  – Accurate estimates of evaluation measures
MTC’s Two Parts

• MTC comprises two separate parts:
  1. An algorithm for selecting documents to judge
  2. A way to evaluate when many judgments are missing

• If you “believe in” one but not the other, you may pick and choose
  – The judgments the algorithm produces can be fed into other evaluation approaches
  – The evaluation approach can be used with judgments from any other method

• They are linked in the algorithm’s stopping condition
MTC Selection Algorithm Outline

• Start with the simplest case: compare two systems by some measure on one topic
• Outline of MTC algorithm:
  – Derive document weights from an algebraic expression of the difference in the measure
  – Order documents by weight and judge the highest-weighted
  – Use the judgment to update the weights
  – Continue until a stopping condition is reached
Detailed Example: Precision

• Say we want to compare two systems by precision at rank k

• First, define the difference in precision:
  – $\Delta\text{prec}@k = \text{prec}_1@k - \text{prec}_2@k$

• Goal: determine sign of $\Delta\text{prec}@k$

• Define $\text{prec}_1@k$, $\text{prec}_2@k$ in terms of relevance:
  
  $\text{prec}_1@k = \frac{1}{k} \sum_{i=1}^{k} \text{rel}_{1,i}$,  \hspace{2mm} \text{prec}_2@k = \frac{1}{k} \sum_{i=1}^{k} \text{rel}_{2,i}$

  – If we knew the values of $\text{rel}_{1,i}$, $\text{rel}_{2,i}$, we would know the sign of $\Delta\text{prec}@k$
Refining $\Delta \text{prec}@k$

- $\text{rel}_{1,i}, \text{rel}_{2,i}$ could represent the same document
  - System 1 places Doc A at rank 1; system 2 places Doc A at rank 4
    $\Rightarrow \text{rel}_{1,1} \equiv \text{rel}_{2,4}$
- No sense in using two different variables to refer to it
  - Number documents independently of their ranking
  - Let $x_i$ indicate the relevance of document number $i$
  - Let $\text{rank}_j(i)$ indicate the rank document $i$ appears at in system $j$
- Now we can write $\Delta \text{prec}@k$ as:

$$\Delta \text{prec}@k = \frac{1}{k} \sum_{i=1}^{n} x_i I(\text{rank}_1(i) \leq k) - \frac{1}{k} \sum_{i=1}^{n} x_i I(\text{rank}_2(i) \leq k)$$

$$= \frac{1}{k} \sum_{i=1}^{n} x_i \left( I(\text{rank}_1(i) \leq k) - I(\text{rank}_2(i) \leq k) \right)$$

- $I(\text{rank}_j(i) \leq k)$ is 1 if document $i$ is ranked above $k$ and 0 otherwise
Precision at rank 5
with local document numbering

System 1

rel₁₁ A
rel₁₂ B
rel₁₂ C
rel₁₂ D
rel₁₂ E

System 2

rel₂₄ G
rel₂₄ E
rel₂₄ C
rel₂₄ A
rel₂₄ H

k = 5
Precision at rank 5 with global document numbering

Document numbers are independent of rank... use rank(i) to map back to rank

\[ \Delta \text{prec@5} = \frac{1}{5} \sum_{i=1}^{n} x_i \left( I(\text{rank}_1(i) \leq 5) - I(\text{rank}_2(i) \leq 5) \right) \]
Goal of MTC

• Decide which subset of $x^n = \{x_1, x_2, \ldots, x_n\}$ to “reveal” (have judged) to prove sign of $\Delta\text{prec}@k$ is -1, 0, or 1

$$\Delta\text{prec}@k = \frac{1}{k} \sum_{i=1}^{n} x_i \left( I(\text{rank}_1(i) \leq k) - I(\text{rank}_2(i) \leq k) \right)$$

• Notice:
  – Judging a document ranked below $k$ by both systems tells us nothing
    • $I(\text{rank}_1(i) \leq k) - I(\text{rank}_2(i) \leq k) = 0 - 0 = 0$
  – Judging a document ranked above $k$ by both systems tells us nothing
    • $I(\text{rank}_1(i) \leq k) - I(\text{rank}_2(i) \leq k) = 1 - 1 = 0$
  – The only interesting documents are those ranked above $k$ by one system but not the other

• Define “interestingness” weight
  – $w_i = I(\text{rank}_1(i) \leq k) - I(\text{rank}_2(i) \leq k)$
Calculating document weights

\[ w_1 = I(\text{rank}_1(1) \leq 5) - I(\text{rank}_2(1) \leq 5) \]
\[ = 1 - 1 \]
\[ = 0 \]

\[ w_2 = 1 \]
\[ w_3 = 1 \]
\[ w_4 = -1 \]
\[ w_5 = 0 \]
\[ w_6 = -1 \]
\[ w_7 = 0 \]
\[ w_8 = 0 \]

Only four documents are useful to judge...
Selecting Documents

• But do we need to judge \textit{all} of the interesting documents?
• After each judgment, ask the following:
  – What is the maximum possible value of $\Delta\text{prec}@k$?
  – What is the minimum possible value of $\Delta\text{prec}@k$?
• Check these values:
  – If the maximum possible is less than zero, then we have proved
    that $\text{sign}(\Delta\text{prec}@k) = -1$; no more judging is necessary
  – If the minimum possible is greater than zero, we have proved
    that $\text{sign}(\Delta\text{prec}@k) = 1$; no more judging is necessary
  – Otherwise we must keep judging
• In other words, bound $\Delta\text{prec}@k$
  – Calculate lower and upper bounds by making different
    assumptions about the relevance of the unjudged documents
Bounding $\Delta$precision@5

Starting from no relevance judgments:

Upper bound: B, D relevant
G, H not relevant

Lower bound: B, D not relevant
G, H relevant

$-0.4 \leq \Delta\text{prec}@ \leq 0.4$

We cannot conclude anything.
Bounding $\Delta\text{precision}@5$

Suppose B and D are judged relevant. Then:

Upper bound: no more relevant G, H not relevant

Lower bound: no more not relevant G, H relevant

$0.0 \leq \Delta\text{prec}@ \leq 0.4$

We conclude that system 2 cannot be better than system 1.

We don’t know whether system 1 is better than system 2.
Bounding $\Delta$precision@$5$

Suppose B and D are judged *non*relevant. Then:

Upper bound: no more relevant
G, H not relevant

Lower bound: no more not relevant
G, H relevant

$-0.4 \leq \Delta prec@ \leq 0.0$

We conclude that system 1 cannot be better than system 2.

We don’t know whether system 2 is better than system 1.

Whether documents judged relevant or not relevant, effect on bounds is the same.
Bounding $\Delta$precision@$k$

- The bounds can be expressed with simple formulas:

$$[\Delta prec@k] = \frac{1}{k} \left( \sum_{i|\text{i judged}} w_i x_i + \#(\text{unjudged and } w_i > 0) \right)$$

$$[\Delta prec@k] = \frac{1}{k} \left( \sum_{i|\text{i judged}} w_i x_i - \#(\text{unjudged and } w_i < 0) \right)$$

Contribution of judged documents

Contribution of unjudged system 1-only documents

Contribution of unjudged system 2-only documents
The Algorithm (MTC for prec@k)

• for each doc $i$ from 1 to $n$,
  – set $w_i = I(\text{rank}_1(i) \leq k) - I(\text{rank}_2(i) \leq k)$
• lowerbound = 0; upperbound = 0
• while (lowerbound $\leq 0$ and upperbound $\geq 0$)
  – Judge an unjudged document with $|w_i| > 0$
    • Alternate between docs with $w_i = 1$, $w_i = -1$
  – Recompute $\Delta \text{prec}@k$ bounds:
    • lowerbound $= \frac{1}{k} \left( \sum_{i| i \text{ judged}} w_i x_i - \#(\text{unjudged and } w_i < 0) \right)$
    • upperbound $= \frac{1}{k} \left( \sum_{i| i \text{ judged}} w_i x_i + \#(\text{unjudged and } w_i > 0) \right)$
MTC is Minimal

• Theorem: MTC requires the minimal number of judgments to determine the sign of $\Delta \text{prec@k}$
  – More precisely: among all algorithms with no prior information about relevance, MTC requires no more judgments on average than any of them
    • Algorithms that learn something about the distribution of relevant documents (such as MTF) could do better
    • MTC could do worse on some cases while still doing better on average
MTC is Minimal: Proof Sketch

• First define two probabilities:
  – $p_1$ is the probability that a document unique to system 1 is judged relevant
    • i.e. the probability that a doc with $w_i > 0$ is relevant
  – $p_2$ is defined likewise for system 2

• If $p_1 > p_2$ then system 1 is better than system 2
  – And vice versa
MTC is Minimal: Proof Sketch

• Suppose w.l.o.g. that $p_1 > p_2$
• Suppose MTC stops after m judgments
  – At this point the lower bound is greater than zero
  – Because of alternation, $m/2$ of the judged documents are from system 1, $m/2$ from system 2
• We can place non-MTC algorithms in one of two bins:
  – Those that might judge documents with $w_i = 0$ (the majority)
  – Those that select among the same set as MTC but do not alternate (“MTC-like”)
MTC is Minimal: Proof Sketch

• Suppose an alternative approach also selects $m$ documents to judge

• If even one of those has $w_i = 0$, then the lower bound of $\Delta \text{prec}@k$ cannot be greater than zero
  – At least one more judgment will be required to complete the proof

• This encompasses all non-MTC-like approaches
MTC is Minimal: Proof Sketch

• For MTC-like approaches, the argument is more difficult
• The idea is as follows:
  – Since an MTC-like approach only judges documents with \( w_i \neq 0 \), the only difference is that it does not alternate between \( w_i > 0 \) and \( w_i < 0 \)
  – This means it prefers documents unique to system 1 or documents unique to system 2
  – Because of this preference, it may be able to prove one bound faster, but it won’t be able to prove the other bound faster
  – Therefore it cannot do better than MTC
MTC for DCG@k

• DCG has become a popular measure due to its use of a user model and graded judgments
  – Gain function $g(x_i)$ maps judgments to gain values
  – Discount function $d(rank(i))$ discounts gains by rank
  – DCG is a family of measures with particular cases defined by specific $g()$ and $d()$

• As we did with precision, define DCG in terms of relevance variables $x_i$ and their ranks $rank(i)$:

$$DCG@k = \sum_{i=1}^{n} \frac{g(x_i)}{d(rank(i))} I(rank(i) \leq k)$$
MTC for DCG@k

• Now we can define the difference $\Delta \text{DCG}@k$:

$$\Delta \text{DCG}@k = \sum_{i=1}^{n} \frac{g(x_i)}{d(\text{rank}_1(i))} I(\text{rank}_1(i) \leq k) - \frac{g(x_i)}{d(\text{rank}_2(i))} I(\text{rank}_2(i) \leq k)$$

$$= \sum_{i=1}^{n} g(x_i) \left( \frac{I(\text{rank}_1(i) \leq k)}{d(\text{rank}_1(i))} - \frac{I(\text{rank}_2(i) \leq k)}{d(\text{rank}_2(i))} \right)$$

• ... and the document weights:

$$w_i = \frac{I(\text{rank}_1(i) \leq k)}{d(\text{rank}_1(i))} - \frac{I(\text{rank}_2(i) \leq k)}{d(\text{rank}_2(i))}$$

• This is similar to precision, but now the ranks matter as well as whether it was retrieved
DCG at rank 5

\[ w_1 = \frac{1}{\log_2(5+1)} - \frac{1}{\log_2(2+1)} = -0.244 \]
\[ w_2 = \frac{1}{\log_2(2+1)} - 0/\log_2(8+1) = 0.631 \]
\[ w_3 = \frac{1}{\log_2(4+1)} - 0/\log_2(5+1) = 0.431 \]
\[ w_4 = -0.387 \]
\[ w_5 = 0.569 \]
\[ w_6 = -1.000 \]
\[ w_7 = 0.000 \]
\[ w_8 = 0.000 \]

\[ \Delta DCG@5 = \sum_{i=1}^{n} (2^{x_i} - 1) \left( \frac{I(rank_1(i) \leq 5)}{\log_2(rank_1(i) + 1)} - \frac{I(rank_2(i) \leq 5)}{\log_2(rank_2(i) + 1)} \right) \]
MTC for DCG@k

• Finally, bounds on $\Delta DCG@k$ are:

$$|\Delta DCG@k| = \sum_{i|i \text{ judged}} w_i g(x_i) + \sum_{i|i \text{ unjudged and } w_i > 0} w_i \text{ max gain}$$

$$|\Delta DCG@k| = \sum_{i|i \text{ judged}} w_i g(x_i) + \sum_{i|i \text{ unjudged and } w_i < 0} w_i \text{ max gain}$$
DCG at rank 5

\[ w_1 = -0.244 \]
\[ w_2 = 0.631 \]
\[ w_3 = 0.431 \]
\[ w_4 = -0.387 \]
\[ w_5 = 0.569 \]
\[ w_6 = -1.000 \]
\[ w_7 = 0.000 \]
\[ w_8 = 0.000 \]

\(-0.631 \leq \Delta DCG@5 \leq 1.631\)

\[
\Delta DCG@5 = \sum_{i=1}^{n} (2^{x_i} - 1) \left( \frac{I(rank_1(i) \leq 5)}{\log_2(rank_1(i) + 1)} - \frac{I(rank_2(i) \leq 5)}{\log_2(rank_2(i) + 9)} \right)
\]
Multiple Topics

• We usually evaluate over more than just one topic
• There are two ways to use an MTC algorithm:
  1. Apply it separately to each topic
     Gives a set of signs of measure differences, e.g. 50 values of sign $(\Delta DCG)$
  2. Apply it to all topics simultaneously
     Gives the sign of the mean difference, e.g. the value of sign$(\Delta DCG)$ averaged over 50 topics
• The second is better:
  – That’s the quantity we’re directly interested in
  – It allows the algorithm to find the topics that are interesting as well as the documents
<table>
<thead>
<tr>
<th>Topic 1</th>
<th></th>
<th>Topic 2</th>
<th></th>
<th>Topic 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>System 1</td>
<td>System 2</td>
<td>System 1</td>
<td>System 2</td>
<td>System 1</td>
<td>System 2</td>
</tr>
<tr>
<td>x_5</td>
<td>A</td>
<td>x_6</td>
<td>G</td>
<td>x_1</td>
<td>A</td>
</tr>
<tr>
<td>x_2</td>
<td>B</td>
<td>x_1</td>
<td>E</td>
<td>x_3</td>
<td>D</td>
</tr>
<tr>
<td>x_8</td>
<td>C</td>
<td>x_8</td>
<td>C</td>
<td>x_5</td>
<td>A</td>
</tr>
<tr>
<td>x_3</td>
<td>D</td>
<td>x_5</td>
<td>A</td>
<td>x_11</td>
<td>D</td>
</tr>
<tr>
<td>x_1</td>
<td>E</td>
<td>x_4</td>
<td>H</td>
<td>x_9</td>
<td>E</td>
</tr>
<tr>
<td>x_7</td>
<td>F</td>
<td>x_3</td>
<td>D</td>
<td>x_10</td>
<td>B</td>
</tr>
<tr>
<td>x_6</td>
<td>G</td>
<td>x_7</td>
<td>F</td>
<td>x_16</td>
<td>C</td>
</tr>
<tr>
<td>x_4</td>
<td>H</td>
<td>x_2</td>
<td>B</td>
<td>x_11</td>
<td>E</td>
</tr>
</tbody>
</table>

Top 6 highest-weighted docs: G (topic 1), A (topic 3), H (topic 3), B (topic 1), B (topic 3), G (topic 3)
Recall Measures

• Note that precision and DCG do not require knowing how many relevant docs there are
  – That is the real challenge for most low-cost methods

• Can MTC work for recall, NDCG, AP, and other such measures?
  – For individual queries, yes: the denominators don’t affect the difference
  – For a set of queries...?
MTC for Recall

• Again, define recall@k in terms of $x_i$ and rank(i):

$$rec@k = \frac{1}{\sum_{i=1}^{n} x_i} \sum_{i=1}^{n} x_i I(rank(i) \leq k)$$

– The denominator is the total number of relevant documents

• Similarly, a difference in recall:

$$\Delta rec@k = \frac{1}{\sum_{i=1}^{n} x_i} \left( \sum_{i=1}^{n} x_i (I(rank_1(i) \leq k) - I(rank_2(i) \leq k) \right)$$

• To define weights, ask: what happens to our understanding of recall when we judge a document?
With no documents judged, what are the max/min values of Δrec@5?

B, D relevant; G, H nonrelevant  \[\Rightarrow \Delta \text{rec@5} = 1.0\]

B, D nonrelevant; G, H relevant  \[\Rightarrow \Delta \text{rec@5} = -1.0\]
Suppose document B is judged relevant
i.e. \( x_2 = 1 \)

What are the max/min values of \( \Delta \text{rec@5} \)?

B relevant
D relevant; G, H nonrelevant
\[ \Rightarrow \Delta \text{rec@5} = 1.0 \]

B relevant
D nonrelevant; G, H relevant
\[ \Rightarrow \Delta \text{rec@5} = 1/3 - 2/3 = -0.333 \]

So \(-0.333 \leq \Delta \text{rec@5} \leq 1.0\)
Suppose document B is judged not relevant
  i.e. $x_2 = 0$

What are the max/min values of $\Delta \text{rec@5}$?

B nonrelevant
D relevant; G, H nonrelevant
  $\Rightarrow \Delta \text{rec@5} = 1.0$

B nonrelevant
D nonrelevant; G, H relevant
  $\Rightarrow \Delta \text{rec@5} = 0/2 - 2/2 = -1.0$

So $-1.0 \leq \Delta \text{rec@5} \leq 1.0$

Judging B nonrelevant accomplishes nothing!
MTC for Recall

• With precision and DCG, judging a document relevant or not relevant didn’t matter
  – Either way, one of the bounds is affected
  – Effect is equal in both cases
• With recall, it does matter
  – A relevant judgment increases the lower bound
  – A nonrelevant judgment does nothing
• Furthermore, each judgment affects the possible effect of future judgments
Finally: MTC for AP

- Average precision presents an additional challenge: relevance judgments interact
  - If the document at rank 1 is relevant, then the contribution of every subsequent relevant document increases
  - If the document at rank 1 is nonrelevant, then the maximum possible contribution of subsequent relevant documents decreases
• Define SP (Sum Precision) as AP*R
  – SP is between 0 and R
• If document A is relevant, its total contribution to SP is as much as 1+1/2+1/3+...
  – Depending on relevance of subsequent docs
• If document A is not relevant, SP cannot be greater than R-1-1/2-1/3-...
• Judgments of nonrelevance can be informative for AP
MTC for AP

• Define AP in terms of $x_i$ and $\text{rank}(i)$ as follows:

$$AP = \frac{1}{\sum_{i=1}^{n} x_i} \sum_{i=1}^{n} x_i \cdot \frac{1}{\text{rank}(i)} \sum_{j=1}^{n} x_j I(\text{rank}(j) \leq \text{rank}(i))$$

• Note that AP sums over all documents
  – Those that were not retrieved should be assumed to appear at rank infinity

• This can be usefully simplified:

$$AP = \frac{1}{\sum_{i=1}^{n} x_i} \sum_{j \leq i} \frac{1}{a_{ij}} x_i x_j, \quad a_{ij} = \min\{\text{rank}(i), \text{rank}(j)\}$$
MTC for AP

• Now define the difference in AP:

\[ \Delta AP = \frac{1}{\sum x_i} \sum_{j \leq i} c_{ij} x_i x_j \]

\[ c_{ij} = \frac{1}{\max\{rank_1(i), rank_1(j)\}} - \frac{1}{\max\{rank_2(i), rank_2(j)\}} \]

• For simplicity, ignore the denominator for now
Assume all documents are nonrelevant.
What happens if we judge one relevant?

\[ x_1: \ SP_1 = \frac{1}{5}, \ SP_2 = \frac{1}{2}, \ \Delta SP = -0.300 \]

\[ x_2: \ SP_1 = \frac{1}{2}, \ SP_1 = \frac{1}{8}, \ \Delta SP = 0.375 \]

\[ x_3: \ \Delta SP = 0.083 \]

\[ x_4: \ \Delta SP = -0.075 \]

\[ x_5: \ \Delta SP = 0.750 \]

\[ x_6: \ \Delta SP = -0.857 \]

\[ x_7: \ \Delta SP = 0.024 \]

\[ x_8: \ \Delta SP = 0.000 \]
Or assume all documents are relevant
What happens if we judge one nonrelevant?

\[ x_1: \quad SP_1 = 1+1+1+1+5/6+6/7+7/8 \]
\[ SP_2 = 1+2/3+3/4+...+7/8 \]
\[ \Delta SP = 0.783 \]

\[ x_2: \quad SP_1 = 1+2/3+3/4+...+7/8 \]
\[ SP_1 = 1+1+1+1+1+1+1 \]
\[ \Delta SP = -1.218 \]

\[ x_3: \quad \Delta SP = -0.367 \]

\[ x_4: \quad \Delta SP = 0.434 \]

\[ x_5: \quad \Delta SP = -1.083 \]

\[ \boxed{x_6: \quad \Delta SP = 1.593} \]

\[ x_7: \quad \Delta SP = -0.143 \]

\[ x_8: \quad \Delta SP = 0.000 \]
Calculating Document Weights

• Initially each document gets a “relevant weight” and a “nonrelevant weight”
  – Relevant weight = effect on \( \Delta SP \) if relevant
    \[ = c_{ii} \]
  – Nonrelevant weight = effect on \( \Delta SP \) if nonrelevant
    \[ = c_{ii} + c_{1i} + c_{2i} + c_{3i} + \ldots + c_{ni} \]

• Judge the document with the greatest maximum of rel weight and nonrel weight
G judged nonrelevant ($x_6 = 0$)
Assume all documents are nonrelevant
What happens if we judge one relevant?

$x_1$: \(SP_1 = 1/5\), \(SP_2 = 1/2\)
\[\Delta SP = -0.300\]

$x_2$: \(SP_1 = 1/2\), \(SP_1 = 1/8\)
\[\Delta SP = 0.375\]

$x_3$: \(\Delta SP = 0.083\)

$x_4$: \(\Delta SP = -0.075\)

$x_5$: \(\Delta SP = 0.750\)

$x_6$: \(\Delta SP = 0.857\)

$x_7$: \(\Delta SP = 0.024\)

$x_8$: \(\Delta SP = 0.000\)
Or assume all documents are relevant
What happens if we judge one nonrelevant?

\[ x_1: \quad SP_1 = 1+1+1+1+5/6+6/8 \]
\[ SP_2 = 1/3+2/4+\ldots+6/8 \]
\[ \Delta SP = 2.019 \]

\[ x_2: \quad SP_1 = 1+2/3+3/4+\ldots+6/8 \]
\[ SP_1 = 1/2+2/3+\ldots+6/7 \]
\[ \Delta SP = 0.393 \]

\[ x_3: \quad \Delta SP = 1.202 \]

\[ x_4: \quad \Delta SP = 1.952 \]

\[ x_5: \quad \Delta SP = 0.402 \]

\[ x_6: \quad \Delta SP = 1.593 \]

\[ x_7: \quad \Delta SP = 1.450 \]

\[ x_8: \quad \Delta SP = 1.402 \]
Updating Document Weights

\[
\begin{align*}
\omega^R_i & = c_{ii} + \sum_{j \mid j \text{ judged}} c_{ij} x_j \\
\omega^N_i & = c_{ii} + \sum_{j \mid j \text{ judged}} c_{ij} x_j + \sum_{j \mid j \text{ not judged}} c_{ij}
\end{align*}
\]

- base effect
- interactions with judged documents
- additional base for nonrel weight
G judged nonrelevant ($x_6 = 0$)
E judged relevant ($x_1 = 1$)

$x_1$: $\Delta SP = 0.600$

$x_2$: $\Delta SP = 0.150$

$x_3$: $\Delta SP = -0.183$

$x_4$: $\Delta SP = -0.450$

$x_5$: $\Delta SP = 0.400$

$x_6$: $\Delta SP = 1.514$

$x_7$: $\Delta SP = 0.252$

$x_8$: $\Delta SP = -0.433$
Stopping Condition

• How do we calculate bounds on ΔSP?
  – A: We don’t. They’re too hard (NP-Hard).
• But we can still determine whether the stopping condition is satisfied
  – “Look ahead”
  – If the algorithm continued in the best case, would our conclusion change?
• If ΔSP > 0 with current judgments, can it become < 0 after a series of future judgments?
So far we know:
  G is nonrelevant
  E is relevant

Based on that, $\Delta SP = -0.3$

Is it possible for system 1 to catch up?

YES: if A is judged relevant,
$\Delta SP$ will go up to 0.4
MTC for AP: Algorithm

• while (!done)
  – for each unjudged document i,
    • $w_i = \max\{w_i^R, w_i^N\}$ (where $w_i^R$, $w_i^N$ calculated as above)
  – judge document with max $|w_i|$ \\
  – calculate $\Delta$AP with current judgments
    • if $\Delta$AP > 0, simulate algorithm forward taking documents in order of increasing $w_i^R$
    • if $\Delta$AP < 0, simulate forward taking documents in order of decreasing $w_i^R$
  – if sign is the same after simulation, done = true
MTC: Summary So Far

• MTC is a family of algorithms with specific cases for each evaluation measure

• An algorithm is defined by
  – A way to weight documents
  – A way to select which document to judge next
  – A way to update document weights
  – A stopping condition based on bounds

• Some algorithms are easier to understand/implement/prove optimal than others...
Refining the Bounds

• Lower and upper bounds are a blunt instrument
  – Bounds can be on the wrong side of zero, but only by a small fraction

• Define a probability distribution over values between the bounds
  – If the total probability of values greater than 0 is low, stop judging
Distributions of Evaluation Measures

• Basic idea:
  – There is a set of \( m \) unjudged documents
  – Each one could be relevant or nonrelevant
  – Thus, there are \( 2^m \) total possible ways to assign relevance to the unjudged documents
  – Each one of those assignments results in a particular value of the measure
  – We can therefore count the number of ways every possible value of \( \Delta \text{prec@k}, \Delta \text{rec@k}, \Delta \text{AP}, \text{etc.} \) can occur
<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
<th>$x_8$</th>
<th>prec$_1@5$</th>
<th>prec$_2@5$</th>
<th>Δprec$_@5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.6</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.4</td>
<td>0.6</td>
<td>-0.2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.0</td>
</tr>
</tbody>
</table>

| | | | | | | | | | | |

<table>
<thead>
<tr>
<th>precision@5</th>
<th>Δprecision@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.0, 0.2]</td>
<td>[0.0, 0.4]</td>
</tr>
<tr>
<td>[0.4, 0.6]</td>
<td>[0.6, 0.8]</td>
</tr>
<tr>
<td>[0.8, 1.0]</td>
<td>[1.0, 1.2]</td>
</tr>
</tbody>
</table>

System 1: $x_5 \rightarrow A \rightarrow x_6$
System 2: $x_5 \rightarrow G \rightarrow x_6$

$k = 5$
Distributions of Evaluation Measures

• Forming a distribution:
  – Assume each of the $2^m$ assignments of relevance is equally likely
    • uniform distribution over possible assignments of relevance
  – Result: values of $\Delta\text{prec@k}$ have a binomial distribution

• As documents are judged, the distribution’s center shifts, but it remains binomial
<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
<th>$x_8$</th>
<th>precision@5</th>
<th>prec2@5</th>
<th>Δprecision@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.6</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.4</td>
<td>0.6</td>
<td>-0.2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>precision@5</th>
<th>Δprecision@5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

System 1

- $x_5$: A
- $x_2$: B
- $x_3$: D
- $x_1$: E
- $x_7$: F
- $x_6$: G
- $x_4$: H

System 2

- $x_5$: A
- $x_1$: B
- $x_8$: C
- $x_3$: D
- $x_7$: F
- $x_6$: G
- $x_4$: H

$k = 5$
Normal Approximations

- The binomial distribution can be approximated by a normal distribution
  - Pretty close approximation even for small k
- It turns out that distributions of $\Delta DCG$ and $\Delta AP$ can also be roughly approximated by normal distributions
  - Proofs possible using combinatoric arguments and limit theory
  - Proofs don’t require uniform distribution of relevance assignments
Using Distributions in MTC

• Since measures are normally distributed, it is very easy to compute the probability that one system will be better than another
  – i.e. given a set of judgments J, we can easily compute $P(\Delta \text{measure} > 0 \mid J)$

• This in turn lets us know whether it’s worth making more judgments
  – Instead of computing bounds, compute a probability
  – If the probability is low, judging can stop
Results

• So how well does MTC actually do?
• Experiment: randomly select a pair of systems, compare them using MTC
  – Validate against “true” results using TREC qrels

<table>
<thead>
<tr>
<th>collection</th>
<th>judgments</th>
<th>% correct</th>
<th>judgments</th>
<th>% correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC-3</td>
<td>367.77</td>
<td>91%</td>
<td>622.04</td>
<td>96%</td>
</tr>
<tr>
<td>TREC-4</td>
<td>411.11</td>
<td>97%</td>
<td>559.44</td>
<td>100%</td>
</tr>
<tr>
<td>TREC-5</td>
<td>408.29</td>
<td>91%</td>
<td>813.76</td>
<td>100%</td>
</tr>
<tr>
<td>TREC-6</td>
<td>354.19</td>
<td>91%</td>
<td>1198.36</td>
<td>96%</td>
</tr>
<tr>
<td>TREC-7</td>
<td>302.59</td>
<td>89%</td>
<td>892.37</td>
<td>93%</td>
</tr>
<tr>
<td>TREC-8</td>
<td>297.44</td>
<td>91%</td>
<td>731.48</td>
<td>100%</td>
</tr>
</tbody>
</table>
Inferring Relevance

• A uniform distribution over relevance assignments is not a good assumption
  – Documents that were not retrieved are as likely to be relevant as documents at rank 1?
• Better estimates of the relevance of individual documents would improve performance
Inferring Relevance

• We want an estimate of the probability that each document is relevant
  – i.e. \( p_i = P(x_i = 1) \)

• Our goal will be to use existing relevance judgments to train a model of relevance

• What we can do that IR systems cannot:
  – Use the judgments for a particular topic as training data, then predict judgments on documents for the same topic
Inferring Relevance

• First assumption we’re going to make:
  – Documents are independently relevant, i.e.
    \[ P(x_1, x_2, ..., x_n) = P(x_1)P(x_2)...P(x_n) \]
  – This is a basic assumption of ad hoc IR and many other IR tasks

• Second assumption:
  – The log of the odds of relevance of a document is a linear combination of feature values
  – This is for simplicity: linear models are easier to fit
Inferring Relevance

• The model is:

\[
\log \frac{p_i}{1 - p_i} = \beta_0 + \sum_{j=1}^{F} \beta_j f_{ij}
\]

– where \( f_{ij} \) is the value of a feature calculated from document \( i \) and \( \beta_j \) is a coefficient

• Note that this is just a logistic regression model, appropriate for binary judgments
  – Graded judgments would require an ordinal model
Features for Inferring Relevance

• Features can be anything appropriate for predicting relevance

• Some we have tried:
  – Document similarity features
  – System performance features
  – Click features

• The following slides will discuss each in slightly more detail
Document Similarities

• Using document similarities as features is inspired by van Rijsbergen’s Cluster Hypothesis:
  – *Closely associated documents tend to be relevant to the same requests*
• Take a shallow pool of documents to be “features”
• Feature values for document $i$ are its similarities to every document in that pool
System Performance

• Use features derived from the systems being evaluated, such as:
  – Number of (known) relevant documents retrieved
  – Ranks at which relevant documents appear
  – Precisions at ranks of relevant documents

• Inspiration is the “metasearch hypothesis” (cf. Joon Ho Lee):
  – Systems tend to retrieve the same relevant documents but different nonrelevant documents
Clicks

• If available, the number of clicks on a document may be indicative of its relevance

• Some complications:
  – Presentation bias: higher ranks are preferred even if less relevant
  – Interactions: relevance of document at rank i can affect clicks at rank j
MTC Evaluation

• As we said earlier, MTC evaluation is separate from its document selection
  – We could use MTC judgments with the usual assumption: that unjudged docs are not relevant
    • Since MTC is not trying to find all the relevant documents, this is probably not appropriate, though
  – We could use bpref or Q-measures that explicitly account for whether a document is judged or not

• MTC evaluation instead uses the idea of forming a distribution over possible values of the evaluation measure
MTC Evaluation

• The idea is the same as with the stopping condition:
  – We used distribution of ΔAP to calculate $P(ΔAP > 0)$
  – Now we will just look at the distribution of $P(ΔP)$
• But a distribution is not an evaluation measure
• For a single-number summary, calculate the expectation of the distribution
MTC Evaluation

- Since we’re assuming documents are independently relevant, expectations are easy

\[
E[\text{prec@k}] = \frac{1}{k} \sum_{i=1}^{n} p_i I(\text{rank}(i) \leq k)
\]

\[
E[R] = \sum_{i=1}^{n} p_i
\]

\[
E[\text{rec@k}] \approx \frac{1}{E[R]} \sum_{i=1}^{n} p_i I(\text{rank}(i) \leq k)
\]

\[
E[\text{AP}] \approx \frac{1}{E[R]} \sum_{i=1}^{n} c_{ii} p_i + \sum_{i<j} c_{ij} p_i p_j
\]
MTC Evaluation

• What we can show:
  – Although E[AP] is an approximation, the error is on the order of $2^{-n}$ in the size of the collection
  – Variance of AP is also computable in $O(n^3)$ time

• What we cannot show:
  – That E[AP] is a good estimate of the actual value of AP
    • In practice it is not: our relevance models tend to overestimate relevance, leading to low values of E[AP]
MTC Evaluation: Example

<table>
<thead>
<tr>
<th>run</th>
<th>topic</th>
<th>eR</th>
<th>eAP</th>
<th>eRprec</th>
<th>eP5</th>
<th>eP10</th>
</tr>
</thead>
<tbody>
<tr>
<td>udelIndDRPR</td>
<td>1</td>
<td>3518.66</td>
<td>0.0177</td>
<td>0.0569</td>
<td>0.0433</td>
<td>0.1681</td>
</tr>
<tr>
<td>udelIndDRSP</td>
<td>1</td>
<td>3518.66</td>
<td>0.0830</td>
<td>0.1129</td>
<td>1.0000</td>
<td>0.9857</td>
</tr>
<tr>
<td>udelIndDMRM</td>
<td>1</td>
<td>3518.66</td>
<td>0.0792</td>
<td>0.1101</td>
<td>1.0000</td>
<td>0.9857</td>
</tr>
</tbody>
</table>

**summary results:**

<table>
<thead>
<tr>
<th>run</th>
<th>eMAP</th>
<th>eRprec</th>
<th>eP5</th>
<th>eP10</th>
</tr>
</thead>
<tbody>
<tr>
<td>udelIndDRPR</td>
<td>0.030971</td>
<td>0.090344</td>
<td>0.265973</td>
<td>0.295068</td>
</tr>
<tr>
<td>udelIndDMRM</td>
<td>0.046869</td>
<td>0.103990</td>
<td>0.231451</td>
<td>0.323774</td>
</tr>
<tr>
<td>udelIndDRSP</td>
<td>0.047082</td>
<td>0.104238</td>
<td>0.277171</td>
<td>0.356119</td>
</tr>
</tbody>
</table>

UDel results from TREC 2009 Web track (ad hoc task)
MTC Evaluation: Example

Summary results:

<table>
<thead>
<tr>
<th>Run</th>
<th>eMAP</th>
<th>eRprec</th>
<th>eP5</th>
<th>eP10</th>
</tr>
</thead>
<tbody>
<tr>
<td>udelIndDRPR</td>
<td>0.030971</td>
<td>0.090344</td>
<td>0.265973</td>
<td>0.295068</td>
</tr>
<tr>
<td>udelIndDMRM</td>
<td>0.046869</td>
<td>0.103990</td>
<td>0.231451</td>
<td>0.323774</td>
</tr>
<tr>
<td>udelIndDRSP</td>
<td>0.047082</td>
<td>0.104238</td>
<td>0.277171</td>
<td>0.356119</td>
</tr>
</tbody>
</table>

Pairwise comparisons:

<table>
<thead>
<tr>
<th>Run</th>
<th>udelIndDMRM</th>
<th>udelIndDRSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>udelIndDRPR</td>
<td>-0.0159</td>
<td>-0.0161</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>udelIndDMRM</td>
<td></td>
<td>-0.0002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5551</td>
</tr>
</tbody>
</table>

UDel results from TREC 2009 Web track (ad hoc task)
MTC in Practice

• Practical considerations include:
  – Selecting documents when more than two systems are involved
    • Simple solution: judge the document with maximum weight across all pairs—computable in linear time
  – Deciding which documents to predict relevance
    • Usually infeasible to do all of them, instead restrict to pool of retrieved documents
  – “Unbiasing” expected evaluation measures
    • Possibly using priors to keep relevance models from overestimating—work in progress
MTC Summary

• MTC is a family of algorithms for selecting documents to judge
  – The probabilistic stopping condition of those algorithms also produces an evaluation measure
• The best way to use MTC is to compare systems
• The best way to interpret it is with the probability that one system is better than another
  – i.e. \( P(\Delta AP > 0) \)
  – This is the quantity that tells you whether you can have confidence that the judgments are sufficient