A Statistical Method for System Evaluation Using Incomplete Judgments

Jay Aslam
Virgil Pavlu
Emine Yilmaz

Northeastern University
The Problem
The Problem

- Large-scale retrieval evaluation is expensive
- 50k to 100k documents assessed per year in TREC
The Problem

- Large-scale retrieval evaluation is expensive
  - 50k to 100k documents assessed per year in TREC
- Q1: How to scale to larger collections?
The Problem

• Large-scale retrieval evaluation is expensive
  • 50k to 100k documents assessed per year in TREC
• Q1: How to scale to larger collections?
• Q2: How do other organizations cope?
The Problem

• Large-scale retrieval evaluation is expensive
  • 50k to 100k documents assessed per year in TREC
• Q1: How to scale to larger collections?
• Q2: How do other organizations cope?
• A: More efficient evaluation techniques...
Our Contributions: Ideas

- View performance measures as the outcome of a random experiment
  - $PC(k)$, $AP$, $RP$, ...

- Apply sampling theory to efficiently estimate this outcome
  - Analogy: forecasting election results

- Many other benefits...
Our Contributions: Methodology
Our Contributions: Methodology

- An accurate and efficient method for estimating measures from a random sample
  - AP, PC(k), RP, R, ...
  - unbiased by statistical design
  - low variance via non-uniform sampling
  - scalable (many lists, many measures, simultaneously)
  - reusable (can estimate quality of unseen lists)
Outline

• A simple example in detail: estimating PC(k)
• Estimating AP
• Results on TREC data
Consider PC(1000)
Consider PC(1000)

- Obvious solution:
  - examine top 1000 documents
  - return \#rel-seen/1000
Consider PC(1000)

• **Obvious solution:**
  • examine top 1000 documents
  • return #rel-seen/1000

• **Alternate solution:**
  • uniformly sample documents from top 1000
  • judge random sample
  • return #rel-judged/sample-size
Carefully Define Random Experiment

• Sample space: top 1000 docs
  • \( i \in \{1, \ldots, 1000\} \)

• Distribution: uniform over top 1000 docs
  • \( p_i = 1/1000 \)

• Random variable: \( X = \text{relevance of doc} \)
  • \( x_i = \text{rel}(i) = \begin{cases} 0 & \text{if } d_i \text{ is non-relevant} \\ 1 & \text{if } d_i \text{ is relevant} \end{cases} \)
Verify Expectation

\[ E[X] = \sum_{i=1}^{1000} p_i \cdot x_i \]

\[ = \frac{1}{1000} \sum_{i=1}^{1000} x_i \]

\[ = \frac{\#\text{rel}}{1000} \]
Law of Large Numbers
Central Limit Theorem

- Average of random sample converges to mean

\[
\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j \to E[X]
\]

- Average of random sample rapidly becomes Gaussian

\[
E[X] = \mu \\
Var[X] = \sigma^2
\]

\[
\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j \to N(\mu, \sigma/\sqrt{n})
\]
Back to PC(1000)

• Suppose that, unknown to us, PC(1000) = p = 0.25
  • i.e., there were 250 rel docs in top 1000

• How many samples until we could estimate that accurately, say +/- 0.03?

• LLN & CLT
LLN and CLT Illustration: N=16

binomial distribution N=16 p=0.25
LLN and CLT Illustration: N=40
LLN and CLT Illustration: N=100

binomial distribution N=100 p=0.25
Normal Distribution: 95% CI

The graph shows a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The 95% confidence interval is indicated by the red lines at $\pm 2\sigma$. This means that 95% of the data falls within $2\sigma$ of the mean.
Sample Size Calculation

\[
\text{Var}[X] = p \cdot (1 - p) = 3/16
\]

\[
\text{Var}[\bar{X}] = (3/16)/n
\]

\[
\sigma = \sqrt{(3/16)/n}
\]

\[
2\sigma = 2\sqrt{(3/16)/n} = 0.03
\]

\[
n \approx 833
\]
Questions and Concerns
Questions and Concerns

• That’s a lot of samples...
Questions and Concerns

• That’s a lot of samples...
• Is there a “smarter” sampling strategy?
  • Yes: importance sampling
Questions and Concerns

• That’s a lot of samples...
• Is there a “smarter” sampling strategy?
  • Yes: importance sampling
• What about multiple lists (multiple measures) with one sample?
  • Yes: scaling factors
PC(1000): Non-uniform Sampling

Q: How to correct for mean?
A: Scaling factors.
New Random Experiment

- Sample space: top 1000 docs
  - \( i \in \{1, \ldots, 1000\} \)
- Distribution: non-uniform
  - \( p_i = \begin{cases} 
  1.5/1000 & 1 \leq i \leq 500 \\
  0.5/1000 & 501 \leq i \leq 1000 
\end{cases} \)
- Random variable: fix expectation (scaling factors)
  - \( x_i = \begin{cases} 
  \text{rel}(i) \cdot 2/3 & 1 \leq i \leq 500 \\
  \text{rel}(i) \cdot 2 & 501 \leq i \leq 1000 
\end{cases} \)

\[ E[X] \text{ as before...} \]
What About Variance?

Consider two cases:

- All 250 relevant docs in top half (good)
- All 250 relevant docs in bottom half (bad)
Case 1: Relevant Docs in Top Half

\[
\text{Var}[X] = E[X^2] - E^2[X] \\
= \sum_{i=1}^{1000} p_i \cdot x_i^2 - (1/4)^2 \\
= \sum_{i=1}^{500} p_i \cdot x_i^2 + \sum_{i=501}^{1000} p_i \cdot x_i^2 - 1/16 \\
= 250 \cdot \frac{3/2}{1000} \cdot (2/3)^2 + 0 - 1/16 \\
= 1/6 - 1/16 \\
\approx 5/48 \\
\approx 0.1042
\]
Case 2: Relevant Docs in Bottom Half

\[
\text{Var}[X] = E[X^2] - E^2[X] \\
= \sum_{i=1}^{1000} p_i \cdot x_i^2 - \frac{1}{4} \\
= \sum_{i=1}^{500} p_i \cdot x_i^2 + \sum_{i=501}^{1000} p_i \cdot x_i^2 - \frac{1}{16} \\
= 0 + 250 \cdot \frac{1}{2} \cdot \frac{1}{1000} \cdot 2^2 + 0 - \frac{1}{16} \\
= \frac{1}{2} - \frac{1}{16} \\
= \frac{7}{16} \\
\approx 0.4375
\]
Comparison

- Original variance: \( \frac{3}{16} = 0.1875 \)
- Case 1 variance: \( \frac{5}{48} = 0.1042 \)
- Case 2 variance: \( \frac{7}{16} = 0.4375 \)
- Effect on sample size required...
  - 833 (original)
  - 463 (Case 1)
  - 1944 (Case 2)
Conclusion
Conclusion

• Sample where the relevant docs are!
  • importance sampling
Conclusion

• Sample where the relevant docs are!
  • importance sampling

• Use scaling factors to correct expectation
  • ratio between “required” and actual distribution
Conclusion

- Sample where the relevant docs are!
  - importance sampling
- Use scaling factors to correct expectation
  - ratio between “required” and actual distribution
- Many benefits...
  - variance reduction
  - many measures from one sample
  - many lists from one sample
  - reusability
PC(k): Sampling for Many Runs

ranked list → ranked list → ranked list

variance reduction → average

Sampling → judged docs (counts)
Estimating PC

Scaling factors

Evaluated docs
(counts)

PC
What About AP?

List:

<table>
<thead>
<tr>
<th>R</th>
<th>N</th>
<th>R</th>
<th>N</th>
<th>N</th>
<th>R</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1/1</td>
<td>2/3</td>
<td>3/6</td>
<td>4/10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
AP = \frac{1 + 2/3 + 3/6 + 4/10}{4} \approx 0.6417
\]
Average Precision

\[
AP = \frac{1}{R} \cdot \sum_{i : \text{rel}(i)=1} PC(i)
\]

\[
= \frac{1}{R} \cdot \sum_{i=1}^{N} \text{rel}(i) \cdot PC(i)
\]

\[
= \frac{1}{R} \cdot \sum_{i=1}^{N} \text{rel}(i) \sum_{j=1}^{i} \text{rel}(j) / i
\]

\[
= \frac{1}{R} \cdot \sum_{1 \leq j \leq i \leq N} \frac{1}{i} \cdot \text{rel}(i) \cdot \text{rel}(j)
\]

We will sample for the summation (and R)
Analogy to PC(1000)

\[ AP = \frac{1}{R} \cdot \sum_{i : rel(i) = 1} PC(i) \]

\[ = \frac{1}{R} \cdot \sum_{i = 1}^{N} rel(i) \cdot PC(i) \]

\[ = \frac{1}{R} \cdot \sum_{i = 1}^{N} rel(i) \sum_{j = 1}^{i} rel(j)/i \]

\[ = \frac{1}{R} \cdot \sum_{1 \leq j \leq i \leq N} \frac{1}{i} \cdot rel(i) \cdot rel(j) \]

\[ = \sum_{i = 1}^{1000} \frac{1}{1000} \cdot rel(i) \]
Analogy to PC(1000)

\[
\text{AP} = \frac{1}{R} \cdot \sum_{i : \text{rel}(i) = 1} \text{PC}(i) = \frac{1}{R} \cdot \text{PC}(1000) = \\
= \frac{1}{R} \cdot \sum_{i=1}^{N} \text{rel}(i) \cdot \text{PC}(i) = \\
= \frac{1}{R} \cdot \sum_{i=1}^{N} \text{rel}(i) \sum_{j=1}^{i} \text{rel}(j)/i = \\
= \frac{1}{R} \cdot \sum_{1 \leq j \leq i \leq N} \frac{1}{i} \cdot \text{rel}(i) \cdot \text{rel}(j) = \sum_{i=1}^{1000} \frac{1}{1000} \cdot \text{rel}(i)
\]
Analogy to PC(1000)

\[
AP = \frac{1}{R} \cdot \sum_{i : \text{rel}(i) = 1} \text{PC}(i) 
\]

\[
= \frac{1}{R} \cdot \sum_{i=1}^{N} \text{rel}(i) \cdot \text{PC}(i)
\]

\[
= \frac{1}{R} \cdot \sum_{i=1}^{N} \text{rel}(i) \sum_{j=1}^{i} \text{rel}(j)/i
\]

\[
= \frac{1}{R} \cdot \sum_{1 \leq j \leq i \leq N} \frac{1}{i} \cdot \text{rel}(i) \cdot \text{rel}(j)
\]

Random Variable

\[
= \sum_{i=1}^{1000} \frac{1}{1000} \cdot \text{rel}(i)
\]
Define the AP Random Experiment

- Sample space:
  - all pairs of docs

- Distribution:
  - as defined above \((1/i)\)

- Random variable:
  - product of relevances

- Many details...
Experiments with TREC8 Data

- 50 queries
- 129 search engine runs per query
- Depth-100 pooling
  - 2/3 runs contributed to pool (training runs)
  - 1/3 did not (testing runs)
- Employ sampling methodology on training runs
  - single sample drawn per query
- Evaluate and compare
  - all runs, all measures
MAP Estimates: 40 JPQ

Sampling MAP, TREC8 K=40

- Training RMS: 0.029417
- Testing RMS: 0.020415
- ρ = 0.976546
- τ = 0.852920
MAP Estimates: 95 JPQ
MPC(30) Estimates: 40 JPQ

Sampling MPC30 TN=8  K=40

trainingRMS=0.028997
 testingRMS=0.023198
 ρ=0.977793
 τ=0.847409
MPC(30) Estimates: 95 JPQ
MAP Summary: RMS

RMS test error for AP estimates TREC8

- **RMS test error**
- **Percentage of pool judged**
- **Depth pooling**
- **Sampling (10 runs avg)**
MAP Summary: Kendall’s tau

Kendall's tau for AP estimates TREC8

- Depth pooling
- Sampling (10 runs avg)

Percentage of pool judged vs. Kendall's tau
Summary and Future Work
Summary and Future Work

- A statistical method for estimating performance
  - AP, RP, PC(k), R, ...
  - view measure as outcome of a random experiment
  - employ sampling theory
  - accurate, efficient estimates
  - reusable
Summary and Future Work

• A statistical method for estimating performance
  • AP, RP, PC(k), R, ...
  • view measure as outcome of a random experiment
  • employ sampling theory
  • accurate, efficient estimates
  • reusable

• Future work:
  • optimal sampling distribution, lower bounds, etc.
  • explicit confidence intervals (estimate variance)
  • mixed strategies (greedy & statistical)
Query Variance: 40 Judgments
Query Variance: 260 Judgments

Sampling AP vs trec AP, all Q system=input.Sab8A1 K=260
Query Variance: 500 Judgments

Sampling AP vs trec AP, all Q, system=input.Sab8A1, K=500
Query Variance: 1000 Judgments

Sampling AP vs trec AP, all Q system=input.Sabb8A1 K=1000
Variance Reduction via Sampling

empirical VAR  SYS=Sab8A1  AP=0.255083  TN=8  QN=46  TRIALS=100

0 0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16 0.18 0.2
sample size
VAR

0 100 200 300 400 500 600 700 800 900 1000
sample size
SD Reduction via Sampling

empirical SD    SYS=Sab8A1 AP=0.255083   TN=8 QN=46 TRIALS=100

SD

sample size

0 100 200 300 400 500 600 700 800 900 1000
Variance Reduction: Sampling

VAR for AVG of N Bernoulli coin tosses  PROB(head)=0.250000
SD Reduction: Sampling

The graph shows the standard deviation (SD) of the estimate for the average of N Bernoulli coin tosses, where the probability of heads is 0.250000. The x-axis represents the number of coin tosses, ranging from 0 to 1000, and the y-axis represents the SD of the estimate, ranging from 0 to 0.45. The line on the graph indicates how the SD decreases as the number of coin tosses increases, demonstrating the reduction in SD with larger samples.
What About AP?

List:

\[
\begin{align*}
\text{R} & \quad 1/1 \\
\text{N} & \quad 2/3 \\
\text{R} & \quad 3/6 \\
\text{R} & \quad 4/10 \\
\text{N} & \\
\text{R} & \\
\text{R} &
\end{align*}
\]

\[
AP = \frac{1 + 2/3 + 3/6 + 4/10}{4} \approx 0.6417
\]
Average Precision

\[ \text{AP} = \frac{1}{R} \cdot \sum_{i : \text{rel}(i) = 1} \text{PC}(i) \]

\[ = \frac{1}{R} \cdot \sum_{i=1}^{Z} \text{rel}(i) \cdot \text{PC}(i) \]

\[ = \frac{1}{R} \cdot \sum_{i=1}^{Z} \text{rel}(i) \sum_{j=1}^{i} \text{rel}(j) / i \]

\[ = \frac{1}{R} \cdot \sum_{1 \leq j \leq i \leq Z} \frac{1}{i} \cdot \text{rel}(i) \cdot \text{rel}(j) \]

We will sample for the summation (and R)...
### AP Sampling Distribution

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1/2</td>
<td>1/2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>1/Z</td>
<td>1/Z</td>
<td>1/Z</td>
<td>...</td>
<td>1/Z</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1/2</td>
<td>1/3</td>
<td>...</td>
<td>1/Z</td>
</tr>
<tr>
<td>2</td>
<td>1/2</td>
<td>1</td>
<td>1/3</td>
<td>...</td>
<td>1/Z</td>
</tr>
<tr>
<td>3</td>
<td>1/3</td>
<td>1/3</td>
<td>2/3</td>
<td>...</td>
<td>1/Z</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>1/Z</td>
<td>1/Z</td>
<td>1/Z</td>
<td>...</td>
<td>2/Z</td>
</tr>
</tbody>
</table>

Normalized, appropriately...
Sampling Strategy

- Ranked list
- Required distribution over pairs of docs
- Marginal distribution over docs \( \sim \frac{3}{2} \)
- Sampling distribution over docs (average)
- Efficiency
- Variance reduction
- Judged docs (counts)
- R
Estimating AP

- Sampling distribution over docs
- Induced distribution over pairs of docs
- Judged docs (counts)
- Judged pairs of docs (counts)
- Scaling factors
- Required distribution over pairs of docs
- Ranked list
- Evaluation
- $\hat{R}$
- $\hat{\text{AP}}$
Estimating PC (and RP)

- sampling distribution over docs
- judged docs (counts)
- ranked list
- required distribution over docs
- scaling factors
  - Evaluation
  - PC
MAP Estimates: 260 JPQ

Sampling MAP, TREC8 K=260

- Training RMS = 0.017856
- Testing RMS = 0.014316
- $\rho = 0.995920$
- $\tau = 0.942816$

Estimated MAP vs. MAP
MPC(30) Estimates: 260 JPQ