Large Scale IR Evaluation

Virgil Pavlu
Large Scale IR Evaluation

- People are not well organized
  - nor consistent with each other
- People are publishing everything
- Today’s Culture, Commerce, Science and Military depend on ability to store and make use of information
- IR field is dedicated to organization of information
  - search is the principal component
Large Scale IR Evaluation
Large Scale IR Evaluation

- Critical for research
- Commercially used for optimization
- Can measure many aspects of returned results
Large Scale IR Evaluation

- Critical for research
- Commercially used for optimization
- Can measure many aspects of returned results
- Text REtrieval Conference
  - queries, collections, search engines, performance
  - many tracks every year
  - judged several 100K documents
Large Scale

- TREC8 judged 86000 documents for 50 queries
  - 86000 \times 2 \text{ minutes} = 1 \text{ year (8 hours per day)}
  - even this is not “complete” (only depth-100)
- corporations may judge even more
Large Scale

- TREC8 judged 86000 documents for 50 queries
  - 86000 X 2 minutes = 1 year (8 hours per day)
  - even this is not “complete” (only depth-100)
- corporations may judge even more
- we want many more queries, say 2000
  - 1 year x 2000/50 = 40 years
  - we cannot afford to judge all documents
Motivation

- attempts to evaluate with incomplete judgements
- 200 docs judged per query (12% of depth-100)
Motivation

- attempts to evaluate with incomplete judgements
- 200 docs judged per query (12% of depth-100)

CONTRIBUTIONS
- AP estimation, not only ranking
- confidence
- reusability +/-
Motivation

- attempts to evaluate with incomplete judgements
- 200 docs judged per query (12% of depth-100)

CONTRIBUTIONS
- AP estimation, not only ranking
- confidence
- reusability +/*
Motivation

- attempts to evaluate with incomplete judgements
- 200 docs judged per query (12% of depth-100)

**CONTRIBUTIONS**
- AP estimation, not only ranking
- confidence
- reusability +/*
Overview

- Sampling intuition
- IR Evaluation
- Sampling w/ replacement
- Sampling w/out replacement
- Confidence
- Million Query Experiment
Sampling Intuition

- say I have 10000 animals
- I want to find percentage of sick animals
- obvious solution:
  - examine all 10000
  - return #sick/10000
Sampling Example

- Alternate solution:
  - uniformly sample animals
  - examine the sampled ones
  - return #sick-seen/#samples

- Distribution: uniform over 10000

- Random variable: $X = \begin{cases} 1 & \text{if sick} \\ 0 & \text{if not} \end{cases}$

\[ p_i = \frac{1}{10000} \]
Estimator Accuracy

Suppose that, unknown to us,
Percentage_sick = q = 0.25

- i.e., there were 2500 sick animals
Estimator Accuracy

Suppose that, unknown to us, Percentage_sick = q = 0.25

- i.e., there were 2500 sick animals

How many samples in order to estimate q accurately, say +/- 0.03?

- in practice: if I take |S| samples, what is the 95% confidence interval?

- CLT: Average of i.i.d. r. v. rapidly becomes Gaussian

  - Mean is preserved
  - Variance decreases linearly in n
CLT

N=16       N=40       N=100
CLT

Distribution of the sampled mean approaches gaussian when sample grows
Normal Distribution 95% CI

normal distribution  $\mu=0$  $\sigma=1$

95% interval: $\pm 2\sigma$
Uniform Sample Size

\[ \text{Var}[X] = q(1 - q) = \frac{3}{16} \]

\[ \text{Var}[\bar{X}] = q(1 - q)/n \]
Uniform Sample Size

\[ \text{Var}[X] = q(1 - q) = 3/16 \]
\[ \text{Var}[\bar{X}] = q(1 - q)/n \]

\[ 2\sigma = 2\sqrt{q(1 - q)/n} = 0.03 \]
\[ n = 4q(1 - q)/0.03^2 \approx 833 \]
Uniform Sample Size

\[ \text{Var}[X] = q(1 - q) = \frac{3}{16} \]

\[ \text{Var}\left[\bar{X}\right] = q(1 - q)/n \]

\[ 2\sigma = 2\sqrt{q(1 - q)/n} = 0.03 \]

\[ n = 4q(1 - q)/0.03^2 \approx 833 \]

- in practice use estimated variance
- that’s 833 samples. Is there a “smarter” sampling strategy?
Importance Sampling

- Sample “more” where sick animals are
  - for example categorize/order them by age:
    - 1–5000 old; 5001–10000 young
Importance Sampling

- Sample “more” where sick animals are
  - for example categorize/order them by age:
    - 1–5000 old;  5001–10000 young

\[ q_i = \begin{cases} 
1.5/10000 & i \leq 5000 \\
0.5/10000 & i > 5000 
\end{cases} \]
Importance Sampling

- Sample “more” where sick animals are
- for example categorize/order them by age:
  - 1-5000 old; 5001-10000 young

- How to correct for estimated mean?
- scaling factors = ratio req.distrib to sample.distrib

\[
x_i = \begin{cases} 
sick(i) \cdot \frac{2}{3} & i \leq 5000 \\
sick(i) \cdot 2 & i > 5000
\end{cases}
\]

\[
\hat{E}[X] = \frac{1}{|K|} \sum_{x \in K} \frac{p(x)}{q(x)} \cdot x
\]
2 extreme cases

- All sick in top half
- All sick in bottom half
2 extreme cases

all sick in top half

\[ Var[X] = E[X^2] - E^2[X] \]

\[
= \sum_{i=1}^{10000} p_i \cdot x_i^2 - (1/4)^2
\]

\[
= \sum_{i=1}^{5000} p_i \cdot x_i^2 + \sum_{i=5001}^{10000} p_i \cdot x_i^2 - 1/16
\]

\[
= 2500 \cdot \frac{3/2}{10000} \cdot (2/3)^2 + 0 - 1/16
\]

\[
= 1/6 - 1/16
\]

\[
= 5/48
\]

\[
\approx 0.1042
\]

sample size = 463

all sick in bottom half
### 2 extreme cases

#### all sick in top half

\[ \text{Var}[X] = E[X^2] - E^2[X] \]

\[
\begin{align*}
\text{sample size} &= 463 \\
10000 &= \sum_{i=1}^{10000} p_i \cdot x_i^2 - (1/4)^2 \\
5000 &= \sum_{i=1}^{5000} p_i \cdot x_i^2 + \sum_{i=5001}^{10000} p_i \cdot x_i^2 - 1/16 \\
&= 2500 \cdot \frac{3/2}{10000} \cdot (2/3)^2 + 0 - 1/16 \\
&= 1/6 - 1/16 \\
&= 5/48 \\
&\approx 0.1042
\end{align*}
\]

#### all sick in bottom half

\[ \text{Var}[X] = E[X^2] - E^2[X] \]

\[
\begin{align*}
\text{sample size} &= 1944 \\
10000 &= \sum_{i=1}^{10000} p_i \cdot x_i^2 - (1/4)^2 \\
5000 &= \sum_{i=1}^{5000} p_i \cdot x_i^2 + \sum_{i=5001}^{10000} p_i \cdot x_i^2 - 1/16 \\
&= 0 + 2500 \cdot \frac{1/2}{10000} \cdot 2^2 + 0 - 1/16 \\
&= 1/2 - 1/16 \\
&= 7/16 \\
&\approx 0.4375
\end{align*}
\]
2 extreme cases

- **all sick in top half**
  
  \[
  Var[X] = E[X^2] - E^2[X]
  \]

  \[
  \begin{align*}
  & \sum_{i=1}^{10000} p_i \cdot x_i^2 - (1/4)^2 \\
  & = \sum_{i=1}^{5000} p_i \cdot x_i^2 + \sum_{i=5001}^{10000} p_i \cdot x_i^2 - 1/16 \\
  & = 2500 \cdot \frac{3/2}{10000} \cdot (2/3)^2 + 0 - 1/16 \\
  & = 1/6 - 1/16 \\
  & = 5/48 \\
  & \approx 0.1042
  \end{align*}
  \]

  sample size = 463

- **all sick in bottom half**

  \[
  Var[X] = E[X^2] - E^2[X]
  \]

  \[
  \begin{align*}
  & \sum_{i=1}^{10000} p_i \cdot x_i^2 - (1/4)^2 \\
  & = \sum_{i=1}^{5000} p_i \cdot x_i^2 + \sum_{i=5001}^{10000} p_i \cdot x_i^2 - 1/16 \\
  & = 0 + 2500 \cdot \frac{1/2}{10000} \cdot 2^2 + 0 - 1/16 \\
  & = 1/2 - 1/16 \\
  & = 7/16 \\
  & \approx 0.4375
  \end{align*}
  \]

  sample size = 1944

- **Comparison:**
  - uniform: sample size required = 833
  - skewed, guessing right: sample size = 463
  - skewed, guessing wrong: sample size = 1944
Sampling
Sampling

- w/ replacement
  - counts, scaling factors
  - sampling & estimation very intuitive
  - not the most efficient
  - some complications with IR setup
Sampling

- **w/ replacement**
  - counts, scaling factors
  - sampling & estimation very intuitive
  - not the most efficient
  - some complications with IR setup
- **w/out replacement**
  - inclusion probabilities replace counts, scaling factors
  - sampling strategy complex
  - estimation not trivial
  - Horwitz-Thompson estimator

\[
\hat{\mu}_{HT} = \frac{1}{N} \sum_{k \in S} \frac{v_k}{\pi_k}
\]
Sampling non-Uniform

we can de-couple sampl.distrib from req.distrib

- estimated mean correct
- use a global sample to estimate many things
Sampling non-Uniform

- we can de-couple sampl.distrib from req.distrib
  - estimated mean correct
  - use a global sample to estimate many things
  - sampling “proportional to size” is ideal: minimizes variance

\[ c = \frac{v_k}{\pi_k} = \frac{\sum_k v_k}{\sum_k \pi_k} = \frac{W}{|S|} \]

\[ \hat{\mu}_{HT} = \frac{1}{N} \sum_{k \in S} \frac{v_k}{\pi_k} = \frac{W}{N} \]
Sampling non-Uniform

- we can de-couple sample distrib from req. distrib
  - estimated mean correct
  - use a global sample to estimate many things
  - sampling “proportional to size” is ideal: minimizes variance

\[
c = \frac{v_k}{\pi_k} = \frac{\sum_k v_k}{\sum_k \pi_k} = \frac{W}{|S|}
\]

\[
\hat{\mu}_{HT} = \frac{1}{N} \sum_{k \in S} \frac{v_k}{\pi_k} = \frac{W}{N}
\]

- in practice use a prior (guess) for values
  - like “sick animals are more likely to be old”
Large Scale IR Evaluation

- Sampling intuition
- IR Evaluation
- Sampling w/ replacement
- Sampling w/out replacement
- Confidence
- Million Query Experiment
IR Evaluation

12 CIKM 2003. New Orleans, Louisiana, USA


CIKM 2003 review
CIKM 2003 highlights. 12th ACM International Conference on Information and Knowledge Management, 3-8 November, New Orleans ... smi.ucd.ie/~rinad/papers/cikm03_rep.html - 22k - Cached - Similar pages

Collaborative Filtering Mailing List Archive: [collab@sims] CFP
ACM CIKM 2003 Call For Papers: 12th International Conference on Information and Knowledge ... caliber papers submitted to CIKM 2003 will be accepted. ... www.pdesigner.net/1996/0697.html - 17k - Cached - Similar pages

TOC
Proceedings of the twelfth international conference on Information and knowledge management citation. 2003, New Orleans, LA, USA November 03-08, 2003 ... portal.acm.org/foc.cfm?id=956863&type=proceeding - Similar pages

[Asis-L] CIKM 2003

[PDF] CIKM 2003
File Format: PDF/Adobe Acrobat - View as HTML

Tutorial on Document Clustering
CIKM 2003 Tutorial. Clustering Large and High-Dimensional Data ... Katya Pelekhov and Daniela Rus;"Using Star Clusters for Filtering", CIKM 2000, (pdf) ... www.csee.umbc.edu/~nicholas/clustering/ - 9k - Cached - Similar pages

Conference on Information and Knowledge Management (CIKM)
CIKM has a strong tradition of workshops devoted to emerging areas of database ... The CIKM 2004 web page, The CIKM 2003 Web Page; The CIKM 2002 Web Page ... www.cikm.org/ - 7k - Cached - Similar pages

## Precision and Recall

<table>
<thead>
<tr>
<th>List:</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>1/1</td>
<td>1/4</td>
</tr>
<tr>
<td>R N</td>
<td>1/2</td>
<td>1/4</td>
</tr>
<tr>
<td>R R</td>
<td>2/3</td>
<td>2/4</td>
</tr>
<tr>
<td>R N</td>
<td>2/4</td>
<td>2/4</td>
</tr>
<tr>
<td>R N</td>
<td>2/5</td>
<td>2/4</td>
</tr>
<tr>
<td>R R</td>
<td>3/6</td>
<td>3/4</td>
</tr>
<tr>
<td>R N</td>
<td>3/7</td>
<td>3/4</td>
</tr>
<tr>
<td>R N</td>
<td>3/8</td>
<td>3/4</td>
</tr>
<tr>
<td>R N</td>
<td>3/9</td>
<td>3/4</td>
</tr>
<tr>
<td>R R</td>
<td>4/10</td>
<td>4/4</td>
</tr>
</tbody>
</table>
## Precision and Recall

<table>
<thead>
<tr>
<th>List:</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>1/1</td>
<td>1/4</td>
</tr>
<tr>
<td>RN</td>
<td>1/2</td>
<td>1/4</td>
</tr>
<tr>
<td>RR</td>
<td>2/3</td>
<td>2/4</td>
</tr>
<tr>
<td>RN</td>
<td>2/4</td>
<td>2/4</td>
</tr>
<tr>
<td>RR</td>
<td>2/5</td>
<td>2/4</td>
</tr>
<tr>
<td>RRR</td>
<td>3/6</td>
<td>3/4</td>
</tr>
<tr>
<td>RR</td>
<td>3/7</td>
<td>3/4</td>
</tr>
<tr>
<td>RRR</td>
<td>3/8</td>
<td>3/4</td>
</tr>
<tr>
<td>RRRR</td>
<td>3/9</td>
<td>3/4</td>
</tr>
<tr>
<td>RRRR</td>
<td>4/10</td>
<td>4/4</td>
</tr>
</tbody>
</table>

The graph shows the precision at recall for different recall levels.
**Average Precision**

- **AP** = average of precisions at relevant ranks
- Use 0 for relevant documents not returned
- **AP** = area under prec-recall curve
- Extremely popular
- We can estimate other measures

**List:**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Precision:**

- 1/1
- 2/3
- 3/6
- 4/10

**Calculation:**

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

Let's rewrite Average Precision

\[ AP = \frac{1}{R} \sum_{rel(d)=1} prec@r(d) = \frac{1}{R} \sum_{d} \frac{\sum_{r(f) \leq r(d)} rel(f)}{r(d)} \cdot rel(d) \cdot \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot rel(d) \cdot rel(f) \]
AP Estimation

\[ AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec}@r(d) \]

\[ = \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f) \]
AP Estimation

\[
AP = \frac{1}{R} \sum_{\text{rel}(d) = 1} \text{prec}@r(d)
\]

\[
= \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f)
\]

w/out replacement
AP Estimation

\[
AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec@r}(d) \\
= \frac{1}{R} \sum_{\text{r}(f) \leq \text{r}(d)} \frac{1}{\text{r}(d)} \cdot \text{rel}(d) \cdot \text{rel}(f)
\]
AP Estimation

\[ AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec}@r(d) \]

\[ = \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f) \]
AP Estimation

\[
AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec@r}(d) \quad \text{w/out replacement}
\]

\[
= \frac{1}{R} \sum_{\text{rel}(d)=1} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f)
\]
AP Estimation

\[ AP = \frac{1}{R} \sum_{\text{rel}(d) = 1} \text{prec}\@ r(d) \]

\[ = \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f) \]
AP Estimation

\[
AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec}\,r(d)
\]

\[
= \frac{1}{R} \sum_{\text{r}(f) \leq \text{r}(d)} \frac{1}{\text{r}(d)} \cdot \text{rel}(d) \cdot \text{rel}(f)
\]
AP Estimation

\[
AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec}@r(d)
\]

\[
= \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f)
\]

w/ replacement
AP Estimation

\[ AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec}@r(d) \]

\[ = \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f) \]

w/ replacement
AP Estimation

\[
AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec}@r(d)
\]

\[
= \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f)
\]

w/ replacement
AP Estimation

\[ AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec@r}(d) \]

\[ = \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f) \quad \text{w/ replacement} \]
AP Estimation

\[ AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec}@r(d) \]

\[ = \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f) \]

w/ replacement
AP Estimation

\[
AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec@r}(d)
\]

\[
= \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f)
\]

w/ replacement
**AP Estimation**

\[
AP = \frac{1}{R} \sum_{\text{rel}(d) = 1} \text{prec@r}(d)
\]

\[
= \frac{1}{R} \sum_{\text{r}(f) \leq \text{r}(d)} \frac{1}{\text{r}(d)} \cdot \text{rel}(d) \cdot \text{rel}(f)
\]

- **MANY systems**
  - same relevance (rel), different rankings (r)
  - estimation has to handle a global sample, not a system-based one
Sampling w/ replacement

\[ AP = \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot rel(d) \cdot rel(f) \]

- R is query constant, lets ignore it for now
  - estimate R globally
- Sample pairs \((d, f)\) with probability \(1/r(d)\)
Sampling w/ replacement

$$AP = \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot rel(d) \cdot rel(f)$$

- R is query constant, lets ignore it for now
  - estimate R globally
- Sample pairs (d,f) with probability 1/r(d)
- Sample documents and from all sampled pairs
  - with a budget of say 200, I can sample little more than 100 pairs...
  - or sample 200 docs and use all 40,000 pairs
Sampling w/ replacement

\[ AP = \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f) \]

sample docs such that pairs appear drawn as required
Sampling w/ replacement

\[
AP = \frac{1}{R} \sum_{r(f) \leq r(d)} \frac{1}{r(d)} \cdot \text{rel}(d) \cdot \text{rel}(f)
\]

- sample docs such that pairs appear drawn as required
- symmetrize, normalize, marginalize

<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>\frac{1}{Z}</td>
<td>\frac{1}{Z}</td>
<td>\frac{1}{Z}</td>
<td>...</td>
<td>\frac{1}{Z}</td>
</tr>
</tbody>
</table>

\[
1 \leq 2 \leq 3 \leq ... \leq Z
\]

\[
1 \leq 2 \leq 3 \leq ... \leq Z
\]

\[
\frac{1}{Z} \leq \frac{1}{Z} \leq \frac{1}{Z} \leq ... \leq \frac{2}{Z}
\]
Docs sample distrib

- After marginalization

\[ W(r) = \frac{1}{2Z} \left( 1 + \frac{1}{r} + \frac{1}{r+1} + \cdots + \frac{1}{Z} \right) \]

- raised at 3/2 for variance reduction

- Average for many systems

  - global sampling distrib over docs
  - what happens with the estimator?
Scaling factors

- If we take sample $K$ according to density $q$ instead of true density $p$

  unbiased \[ \hat{E}[X] = \frac{1}{|K|} \sum_{x \in K} \frac{p(x)}{q(x)} \cdot x \]
Scaling factors

- If we take sample $K$ according to density $q$ instead of true density $p$

\[
\hat{E}[X] = \frac{1}{|K|} \sum_{x \in K} \frac{p(x)}{q(x)} \cdot x
\]

- unbiased

- For our setup, our units are pairs of docs

\[
SF = \frac{\text{required pair distribution}}{\text{induced pair distribution}} = \frac{J(d, f)}{I(d, f)}
\]
Scaling factors

- If we take sample $K$ according to density $q$ instead of true density $p$
  
  unbiased
  
  $E[X] = \frac{1}{|K|} \sum_{x \in K} \frac{p(x)}{q(x)} \cdot x$

- For our setup, our units are pairs of docs

  $SF = \frac{\text{required pair distribution}}{\text{induced pair distribution}} = \frac{J(d, f)}{I(d, f)}$

- Estimator becomes

  $SP_s = |s| \cdot \frac{1}{|K_s|} \sum_{(d,f) \in K_s} \text{rel}(d) \cdot \text{rel}(f) \cdot SF_s(d, f)$
sampling w/ replacement

- sampling distribution over docs
- sample docs (counts)
- judged pairs of docs (counts)
- scaling factors
- Evaluation
- ranked list
- required distribution over pairs of docs
- AP
w/ replacement

- **advantages**
  - simple, intuitive
  - sampling and evaluation straightforward
  - good results

- **disadvantages**
  - not efficient
  - subsampling: requires the entire sampling distribution available at evaluation stage
Sampling w/out replacement

\[ AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec@r}(d) \]

w/out replacement
Sampling w/out replacement

\[ AP = \frac{1}{R} \sum_{\text{rel}(d)=1} \text{prec@r}(d) \]

- population units = prec at relevant ranks
  - values are far from "roughly equal"
- sample w/out replacement
  - no counts
- non-uniform
  - no scaling factors, use inclusion probabilities
- need a sampling design
- real-world sampling is done w/out replacement
sampling w/out replacement

prior, sampling and estimation independent
w/out replacement evaluation

• without replacement
  - $\pi_k$ = inclusion probabilities
  - stratified sampling
    • imagine using sequential sampling

• use a ratio estimator
  - estimate precision@rank
    - numerator: HT for sum-precision
    - denominator: HT for $R$

$$\hat{AP} = \frac{\sum_{k \in S} p_k / \pi_k}{\sum_{k \in S} 1 / \pi_k}$$
Stratified sampling

- non-uniform distribution; sample size = 14
Stratified sampling

- non-uniform distribution; sample size = 14
- partition docs in buckets of size 14 each
Stratified sampling

- non-uniform distribution; sample size = 14
- partition docs in buckets of size 14 each
- sample the buckets with replacement 14 times
  - based on the cumulative weight for each bucket
Stratified sampling

- non-uniform distribution; sample size = 14
- partition docs in buckets of size 14 each
- sample the buckets with replacement 14 times
  - based on the cumulative weight for each bucket
- for each bucket, if picked k times, sample uniform without replacement k docs in it
Sampling results: all queries

Sampling MAP, TREC8 $K=29$

- Estimated MAP
- Training RMS = 0.033
- Testing RMS = 0.039
- $\rho = 0.982$
- $\tau = 0.851$

Sampling MAP, TREC8 $K=200$

- Estimated MAP
- Training RMS = 0.011
- Testing RMS = 0.011
- $\rho = 0.998$
- $\tau = 0.957$
Sampling results: one query

- Variance decreases linearly with the number of queries.

![Graph showing the relationship between estimated AP and AP with training RMS, testing RMS, ρ, and τ values.]
sampling results: one system

Sampling AP, TREC8, all Q, run=Sab8A1, K=200

RMS=0.073870
ρ=0.933765
τ=0.800816
Add deterministic judgments

**Sampling (K=71)**

- Training RMS: 0.041393
- Testing RMS: 0.042806
- $p = 0.952746$
- $\tau = 0.854441$

**Depth pooling (K=71)**

- Training RMS: 0.118704
- Testing RMS: 0.133814
- $p = 0.946775$
- $\tau = 0.838545$
Add deterministic judgments

- **Sampling (K=71)**
  - Training RMS = 0.041393
  - Testing RMS = 0.042806
  - $\rho = 0.952746$
  - $\tau = 0.854441$

- **Combined (K=127)**
  - Training RMS = 0.056034
  - Testing RMS = 0.048935
  - $\rho = 0.971088$
  - $\tau = 0.919535$

- **Depth Pooling (K=71)**
  - Training RMS = 0.118704
  - Testing RMS = 0.133814
  - $\rho = 0.946775$
  - $\tau = 0.838545$

- Depth pooling AP estimates, TREC8, Q40, K=71
  - Training RMS = 0.118704
  - Testing RMS = 0.133814
  - $\rho = 0.946775$
  - $\tau = 0.838545$
Sampling results: trends(RMS)
Sampling results: trends (RMS)

RMS = std

sampling CI = ±0.04
depth-pool CI = ±0.14
sampling results: trends(τ)
Confidence intervals

• pair-inclusion probab must be computed
  - difficult for most sampling designs
    \[ \pi_{df} = \begin{cases} 
    \pi_d \cdot \pi_f & \text{if } \pi_d = \pi_f \\
    \frac{|S|-1}{|S|} \pi_d \cdot \pi_f & \text{if } \pi_d \neq \pi_f 
    \end{cases} \]

• estimate variance of the estimator
  - account for documents with 0 inclusion probability
Confidence intervals - TREC8

TREC 8  K=20

RMSE=0.022  
\( \rho=0.974 \)  
\( \tau=0.829 \)

TREC 8  K=100

RMSE=0.022  
\( \rho=0.992 \)  
\( \tau=0.901 \)
Conf intervals

- Robust Track 05
  - 50 known hard queries
- sab05ror1 system
  - most unique rel docs
  - if not sampled, massively under-evaluated
- sampling can warn of miss-evaluation
Sampling w/ replacement

- **advantages**
  - use [deterministic] additional judgments
  - inclusion probab immune to subsampling
  - needs only information about judged documents
  - completely modular, more general

- **disadvantages**
  - sampling design complex
  - inclusion probabilities can be very hard to compute

NEXT: MQ track
Large Scale IR Evaluation

- Sampling intuition
- IR Evaluation
- Sampling w/ replacement
- Sampling w/out replacement
- Confidence

- Million Query Experiment
Million Query Track 07

- Not a million, but 10000 topics
  - orders of magnitude larger than before (50)
- 1800 topics judged for relevance
  - only possible using sampling or MTC
- MTC - greedy selects most resolving docs
  - proposed by Ben Carterette at UMASS Amherst
  - MTC differentiate systems faster
  - but cannot reliably estimate the actual AP
Million Query Track 07
For 149 topics, we had judgments from previous year (MAP TB)
For 149 topics, we had judgments from previous year (MAP TB)

MTC estimates are on a different scale
Million Query Track 07

- ANOVA (Evan Kanoulas)
  - study possible due to large number of queries

- Budgeting: how many queries at what level of judgment is optimal

![Graph showing cost (assessor time) to reach τ=0.9 ranking against number of judgments per query.](image)
Conclusion

- we proposed two sampling technologies
- very efficient
- what if a better measure comes along
  - principle remain; it is going to need to be estimated
  - and likely sampling will work
Contributions and impact

- very efficient
  - theoretically sound
  - efficient implementations
  - used in MQ Track 07
  - some form of sampling used in other tracks

- can run studies like ANOVA because more queries are judged
  - impact on budget

- confidence - a first
  - impact on reusability
Thank You