TAR 3.0 and Training of Predictive Coding Systems

Bill Dimm
December 10, 2015
Topics

- How machines learn – pattern detection
- Tips on reviewing docs for training
- Performance measures: precision & recall
- TAR 1.0: a baseline
  - How to use control sets
  - Random vs. non-random training
- TAR 2.0: better efficiency, especially for low prevalence
- TAR 3.0: good if you don't need to review everything you'll produce
Identifying Patterns
Sorting by Relevance Score
Benefits of Predictive Coding

- Reduce Cost
  - Less human doc review
- See Relevant Docs Earlier
  - Decide to settle before spending too much on e-discovery
- Quality Assurance
  - Detect inconsistent tagging
Doc Review for Training

- Think about how the computer will interpret the tags you apply
- Don't tag doc as non-relevant just because it is duplicative
- Emails with attachments
  - Want to produce the whole family but only part may be relevant
Garbage In, Garbage Out?

- A low threshold for relevance may be better than a very precise view
- System may identify possible mistakes
- Large amount of low-quality training data may be better than a small amount of high-quality
Terminology

• Prevalence (a.k.a. Richness)
  - Percentage of all docs that are relevant

• Recall
  - Percentage of relevant docs found
  - Important for defensibility

• Precision
  - Percentage of docs predicted to be relevant that actually are relevant
  - Important for cost - reduce review
Precision-Recall Curve

\[ R = \frac{6}{10} \]
\[ P = \frac{6}{8} \]
\[ R = \frac{8}{10} \]
\[ P = \frac{8}{16} \]
1) Review training docs
2) Use control set to determine whether training should end
   Back to (1) if additional training is worthwhile
3) Sort remaining docs by relevance score for review/production
4) Sample/test to ensure sufficient recall

- Training could be with random or non-random docs
  - These options are very different!
Control Set
How Much Training Data?

- Training set size should never involve phrases like:
  - 95% confidence
  - +/- 2%
  - Statistically significant sample (this isn't even a thing!)
- Those phrases are about determining how many docs are relevant.
- Training is about which docs are relevant.
Boxes of Gold/Lead: All Same

- 1 million identical boxes. Some contain an ounce of gold, others an ounce of lead.
- Sample 400 random boxes, find that 80 contain gold
  - 20% +/- 5% contain gold with 95% confidence (really 16% to 24%)
- Sample 1,600 random boxes, 320 contain gold
  - 20% +/- 2.5% contain gold with 95% confidence
- Which boxes contain gold? No idea!
Boxes of Gold/Lead: Colored

- Boxes come in different colors. All boxes with the same color contain the same metal.
- How many boxes do we sample to find the gold?
  - Depends on how many colors there are
Amount of Training Depends on Task
Optimal Training

- More training gives better predictions, so fewer non-relevant docs to review (higher precision)
- Benefit of additional training diminishes until not worthwhile
- \( n = \rho N R / P \)
  - \( n \) = number of docs to review
  - \( \rho \) = prevalence
  - \( N \) = number of non-training docs
  - \( R \) = desired recall
  - \( P \) = precision at desired recall
- Do **not** use \( F_1 \) to measure training progress. Use precision at desired recall.
Optimal Training

![Graph showing the relationship between training docs and review docs, indicating the optimal training point.](image-url)
Static Control Set

- Random set of documents reviewed at beginning
- Fails to account for shifting understanding of relevance as review progresses
Rolling Control Set

- If training with random docs, use most recent docs as control set
- Fixes the shifting understanding of relevance problem
How to Select Training Docs

• Representative / Random
• Judgmental
  - Human chooses docs considered to be good examples, e.g., keyword search
• Active Learning
  - Computer chooses docs based on a strategy intended to aid learning
Active Learning Approaches

- Common Outside of E-Discovery
  - Docs estimated to have 50% chance of relevance
  - Details may be specific to classification algorithm
  - Remember the control set animation – points close to the separating boundary had most impact
- Continuous Active Learning (TAR 2.0)
  - Docs most likely to be relevant
  - You were (probably) going to review them anyway
Random vs Non-Random Training

• Random
  - Theoretically sound (sort of) – training docs look like the population

• Judgmental / Active
  - System sees larger number of relevant docs, which is good
  - Bias
  - Probabilities are distorted – relevance scores hard to interpret

• Thought experiment - teaching a child to recognize dogs
  - Do they need to see 9 birds for every dog?
  - How about 99 birds for every dog?
  - Plausibly more efficient: Lots of dogs, a few birds, and some wolves and foxes
Catch & Release Control Set

- Static control set is flawed for non-random training
  - Docs in control set aren't similar to the unreviewed docs
- Identify random control set docs (from full population), then put them back into the review
Higher Dimensions
TAR 1.0 with Random Training
TAR 2.0 (CAL)

1) Review very small set of training docs (single relevant doc is enough)
2) Update predictions and sort remaining docs by relevance score
3) Review small number of docs with highest relevance scores
   Back to (2) unless not many relevant docs left
4) Sample/test to ensure sufficient recall

- System continues to learn throughout.
- No separation between training and review.
- Huge number of relevant training docs.
TAR 2.0

Training + Review (35% recall)

Training + Review (75% recall)
TAR 1.0 v 2.0 Review to Get R=75%
TAR 3.0 (CAL on Cluster Centers)

1) Form conceptual clusters (narrow focus, fixed radius, agglomerative)
2) Review very small set of training docs (single relevant doc is enough)
3) Update predictions for cluster centers and sort by relevance score
4) Review small number of cluster centers with highest relevance scores
   Back to (3) unless not many relevant cluster centers left
5) Generate predictions for full population, then you have a choice:
   A) Produce docs without review (unless potentially privileged)
   B) Produce docs with high relevance scores without review, and perform standard CAL
      on remainder (review top docs, update predictions, and iterate)
   C) Review all docs that are candidates for production using standard CAL
6) Sample/test to ensure sufficient recall

- Training and review are separate, like TAR 1.0
- No control set needed
- Option to produce docs without review if desired
- Free prevalence estimate
TAR 3.0

Training

Review
Prevalence Estimation with TAR 3.0

- Count documents in clusters where the center is relevant
- This is stratified sampling with two wild assumptions
  - Center of cluster has same probability of being relevant as other docs in cluster
  - Clusters that aren't hit have negligible number of relevant docs
- Not statistically justifiable, but works well if clustering is good
TAR 3.0 Prevalence vs. Random
## Workflow Comparison

<table>
<thead>
<tr>
<th></th>
<th>TAR 1.0</th>
<th>TAR 2.0</th>
<th>TAR 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Works for Low Prevalence</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Avoids Control Set</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Early Prevalence Estimate</td>
<td>Yes?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Produce Without Review</td>
<td>Yes</td>
<td>No?</td>
<td>Yes</td>
</tr>
<tr>
<td>See Relevant Docs Early</td>
<td>No?</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Diverse Early View of Relevance</td>
<td>Yes</td>
<td>No?</td>
<td>Yes</td>
</tr>
<tr>
<td>Add Docs to Population Later</td>
<td>Hard</td>
<td>Easy</td>
<td>Easy</td>
</tr>
<tr>
<td>Efficiency - Review Pos. Pred.</td>
<td>Low</td>
<td>High</td>
<td>Medium</td>
</tr>
<tr>
<td>Efficiency - Don't Rev. Pos. Pred.</td>
<td>Medium</td>
<td>?</td>
<td>High</td>
</tr>
</tbody>
</table>