UF Law E-Discovery Distinguished Speaker Series

Technology-Assisted Review: Fact or Fiction?

February 20, 2017

Maura R. Grossman
maura.grossman@uwaterloo.ca

Gordon V. Cormack
gvcormac@uwaterloo.ca

University of Waterloo
Scouts or Analytics?
Drivers or Autonomous Cars?
But, on occasion . . .
Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey
University of Pennsylvania

Research shows that evidence-based algorithms more accurately predict the future than do human forecasters. Yet when forecasters are deciding whether to use a human forecaster or a statistical algorithm, they often choose the human forecaster. This phenomenon, which we call algorithm aversion, is costly, and it is important to understand its causes. We show that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake. In 5 studies, participants either saw an algorithm make forecasts, a human make forecasts, both, or neither. They then decided whether to tie their incentives to the future predictions of the algorithm or the human. Participants who saw the algorithm perform were less confident in it, and less likely to choose it over an inferior human forecaster. This was true even among those who saw the algorithm outperform the human.
Attorneys vs. Algorithms?

https://e-discoveryteam.com/2016/02/24/why-the-google-car-has-no-place-in-legal-search

Why the ‘Google Car’ Has No Place in Legal Search

Hybrid Multimodal is the preferred legal search method of the e-Discovery Team. The Hybrid part of our method means that our Computer Assisted Review, our CAR, uses active machine learning (predictive coding), but still has a human driver. They work together. Our review method is thus like the Tesla’s Model S car with full autopilot capabilities. It is designed to be driven by both Man and Machine. Our CAR is unlike the Google car, which can only be driven by a machine. When it comes to legal document review, we oppose fully autonomous driving. In our view there is no place for a Google car in legal search.
Technology-Assisted Review (“TAR”)

Mission

• Find *substantially all* relevant documents

Approach

• Human provides hint(s) to System
• System presents document(s) for review
• Human codes document(s) for relevance
• Coding provided to System
• Repeat until *mission accomplished!*
The Standard Treatment

Approach

- Human composes Boolean queries
- Human codes retrieved documents
- *Mission accomplished!*
Is TAR Superior to Standard Treatment?

Amenable to scientific inquiry, but …

- What precisely is “TAR”?
- What precisely is “standard treatment”?
- What precisely is meant by “superior”?
- How can superiority be observed?
The Science of Information Retrieval

TREC: The Text REtrieval Conference
National Institute for Standards and Technology (“NIST”)

Legal Track 2006 – 2011
Total Recall Track 2015 – 2016
Humans vs. (2 Kinds of) TAR


By Maura R. Grossman* & Gordon V. Cormack† **

Humans vs. (2 Kinds of) TAR (Cont’d)

Gordon V. Cormack
University of Waterloo
gvcormac@uwaterloo.ca

Maura R. Grossman*
Wachtell, Lipton, Rosen & Katz
mrgrossman@wlrk.com

SIGIR ’14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07.
http://dx.doi.org/10.1145/2600428.2609601.
SIGIR ’14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07.
http://dx.doi.org/10.1145/2600428.2609601.
Obviating the TAR Whisperer

Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review

GORDON V. CORMACK, University of Waterloo
MAURA R. GROSSMAN, Wachtell, Lipton, Rosen & Katz*

We enhance the autonomy of the continuous active learning method shown by Cormack and Grossman (SIGIR 2014) to be effective for technology-assisted review, in which documents from a collection are retrieved and reviewed, using relevance feedback, until substantially all of the relevant documents have been reviewed. Autonomy is enhanced through the elimination of topic-specific and dataset-specific tuning parameters, so that the sole input required by the user is, at the outset, a short query, topic description, or single relevant document; and, throughout the review, ongoing relevance assessments of the retrieved documents. We show that our enhancements consistently yield superior results to Cormack and Grossman’s version of continuous active learning, and other methods, not only on average, but on the vast majority of topics from four separate sets of tasks: the legal datasets examined by Cormack and Grossman, the Reuters RCV1-v2 subject categories, the TREC 6 AdHoc task, and the construction of the TREC 2002 filtering test collection.
Overall Recall vs. Important Documents Only

Overall Recall

Important Recall
Overall Recall vs. Subtopic Recall

Overall Recall

Subtopic Recall
Recall vs. Effort (Documents Reviewed) for TREC 2016 Total Recall Track -- Athome Task.

- **Run**: BMI, Knee, Target
- **Recall**: .945, .950, .926
- **Effort**: 2,452, 3,091, 25,752

Effort for BMI, Knee, Target.
Recall for BMI, Knee, Target.
## Roger and Me

### Kaine Email Project @ LVA

<table>
<thead>
<tr>
<th>Topic</th>
<th>Recall</th>
<th>Precision</th>
<th>F₁</th>
<th>Effort</th>
<th>Recall</th>
<th>Precision</th>
<th>F₁</th>
<th>Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legal Hold</td>
<td>0.97</td>
<td>0.91</td>
<td>0.94</td>
<td>401,960</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>40,522</td>
</tr>
<tr>
<td>Archival</td>
<td>0.89</td>
<td>0.84</td>
<td>0.86</td>
<td>381,819</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
<td>332,410</td>
</tr>
<tr>
<td>Restricted</td>
<td>0.98</td>
<td>0.75</td>
<td>0.84</td>
<td>146,594</td>
<td>0.95</td>
<td>0.80</td>
<td>0.87</td>
<td>38,048</td>
</tr>
</tbody>
</table>
2011 JOLT vs. 2017 Kaine Results

**Similarities** *(accord Roitblat et al. 2010)*

- Recall: Too close to call
- Precision: TAR superior
- Efficiency: TAR superior

**Differences**

- Uncontrolled human input vs. AutoTAR
- Sampled vs. exhaustive manual review
- Volunteer vs. officially rendered coding
- Open vs. blind relevance adjudication
Wait, There’s More!

- Many more datasets and applications
- Important vs. merely relevant documents
- Heterogeneous document types
- Heterogeneous subject matter and subtopics
- Impact of human coding errors on TAR
- Influences on human coding decisions
- Multi-language review
- Real-time filtering
- New TREC 2017 Common Core Track
  github.com/trec-core/2017
Online AutoTAR Demos

• Jeb Bush Email Collection (~290,000)
  cormack.uwaterloo.ca/cal/

• Tim Kaine Email Collection (~140,000)
  cormack.uwaterloo.ca/kaine/