Information Retrieval for E-Discovery

Douglas W. Oard

College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park

Joint work with: Mossaab Bagdouri, Jason Baron, David Doermann, Gordon Cormack, Sergey Golitsynskiy, Maura Grossman, Bruce Hedin, David Kirsch, David Lewis, Falk Scholer, Ian Soboroff, Paul Thompson, Stephen Tomlinson, Jyothi Vinjumur, William Webber
Early E-Discovery

Clinton White House

32 million emails

National Archives

search request

25 persons

for 6 months …

80,000

32 million

200,000

Tobacco Policy
Federal Rules of Civil Procedure

Rule 26(f): At the parties’ planning meeting, issues expected to be discussed include:

– “Any issues relating to disclosure or discovery of **electronically stored information**, including the form or forms in which it should be produced”

– “Any issues relating to preserving discoverable information”
Inside Yesterday’s Black Box

Case Knowledge

Unprocessed Documents

Coded Documents
Inside Today’s Black Box

Case Knowledge
Keyword Search & Linear Review

Unprocessed Documents

“Reasoning”

“Representation”

“Interaction”

Coded Documents
Judge Grimm, writing for the U.S. District Court for the District of Maryland

“all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search”

Inside Tomorrow’s Black Box

Case Knowledge — Technology Assisted Review

Unprocessed Documents — Coded Documents

“Reasoning” — “Representation” — “Interaction”
A Process View

Electronic Discovery Reference Model

Information Management → Identification → Preservation → Collection → Processing → Review → Analysis → Production → Presentation

VOLUME

RELEVANCE

Electronic Discovery Reference Model / © 2009 / v2.0 / edrm.net
What Does “Better” Mean?

INCREASING EFFORT
(time, resources expended, etc.)

INCREASING SUCCESS
(finding relevant documents)

A

B

C

D

“Better” Technique

“Baseline” Technique

INCREASING EFFORT
(time, resources expended, etc.)
## Outside Our Experience Base

<table>
<thead>
<tr>
<th></th>
<th>Web Search</th>
<th>E-Discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>Duration</td>
<td>10 seconds</td>
<td>10 weeks</td>
</tr>
<tr>
<td>Budget</td>
<td>$0.001</td>
<td>$1,000,000</td>
</tr>
<tr>
<td>Goal</td>
<td>High precision</td>
<td>High recall</td>
</tr>
<tr>
<td>Relevance</td>
<td>Personal</td>
<td>Authoritative</td>
</tr>
<tr>
<td>Stopping criterion</td>
<td>Satisfaction</td>
<td>Reasonableness</td>
</tr>
</tbody>
</table>
An IR-Centric Process Model

1. **Formulation**
   - Production request

2. **Acquisition**
   - Collection

3. **Review for Relevance**
   - Responsive ESI

4. **Review for Privilege**
   - Production

5. **Sense-making**
   - Insight
12. On January 1, 2002, Echinoderm announced record results for the prior year, primarily attributed to strong demand growth in overseas markets, particularly China, for its products. The announcement also touted the fact that Echinoderm was unique among U.S. tobacco companies in that it had seen no decline in domestic sales during the prior three years.

13. Unbeknownst to shareholders at the time of the January 1, 2002 announcement, defendants had failed to disclose the following facts which they knew at the time, or should have known:
   a. The Company's success in overseas markets resulted in large part from bribes paid to foreign government officials to gain access to their respective markets;
   b. The Company knew that this conduct was in violation of the Foreign Corrupt Practices Act and therefore was likely to result in enormous fines and penalties;
   c. The Company intentionally misrepresented that its success in overseas markets was due to superior marketing.
   d. Domestic demand for the Company's products was dependent on pervasive and ubiquitous advertising, including outdoor, transit, point of sale and counter top displays of the Company's products, in key markets. Such advertising violated the marketing and advertising restrictions to which the Company was subject as a party to the Attorneys General Master Settlement Agreement ("MSA").
   e. The Company knew that it could be ordered at any time to cease and desist from advertising practices that were not in compliance with the MSA and that the inability to continue such practices would likely have a material impact on domestic demand for its products. …

All documents which describe, refer to, report on, or mention any “in-store,” “on-counter,” “point of sale,” or other retail marketing campaigns for cigarettes.
Estimating Retrieval Effectiveness

Sampling rate = 6/10
Each Rel counts 10/6

Sampling rate = 3/10
Each Rel counts 10/3

\[
\frac{4}{6} \approx 67\% \text{ relevant in this region}
\]

\[
\frac{1}{3} \approx 33\% \text{ relevant in this region}
\]

\[
estRel(S) = \sum_{d \in \text{JudgedRel}(S)} \frac{1}{p(d)}
\]
26 topics

24 topics

2008 Ad Hoc task
Est. Boolean Recall: All Relevant

26 topics

2008 Ad Hoc task
Est. Boolean Recall: Highly Relevant

2008 Ad Hoc task
Inter-Annotator Agreement

<table>
<thead>
<tr>
<th>A1</th>
<th>A2</th>
<th>Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>R</td>
<td>72</td>
</tr>
<tr>
<td>R</td>
<td>N</td>
<td>32</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>27</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>510</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>641</strong></td>
</tr>
</tbody>
</table>

Positive Overlap = 0.550

2010 Interactive task, Enron email families, Topic 301
Evaluation Design

Interactive Task
TREC Interactive Task

Create Complaint & Production Requests (Topics)

Team-TA Interaction & Classifier Training

Sampling & First-Pass Assessment

Appeal & Adjudicate

Analysis & Reporting

Coordinators

Teams

Assessors

Coordinators

Teams

TA

TA

TA

TA

TA

TA

TA
<table>
<thead>
<tr>
<th>UB</th>
<th>Cl</th>
<th>H5</th>
<th>Pitt</th>
<th>AdHoc</th>
<th>N</th>
<th>n</th>
<th>a</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>5,727</td>
<td>46</td>
<td>46</td>
<td>38</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>46</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>11,965</td>
<td>98</td>
<td>98</td>
<td>78</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>995</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>R</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>131</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>R</td>
<td>N</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>1,547</td>
<td>13</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>220</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>1,901</td>
<td>15</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>46</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>17,082</td>
<td>145</td>
<td>145</td>
<td>111</td>
</tr>
<tr>
<td>R</td>
<td>N</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>10,291</td>
<td>84</td>
<td>84</td>
<td>61</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>176</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>N</td>
<td>19</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>7,679</td>
<td>62</td>
<td>61</td>
<td>23</td>
</tr>
<tr>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>9,531</td>
<td>77</td>
<td>77</td>
<td>17</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>8,068</td>
<td>65</td>
<td>65</td>
<td>49</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>101</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>73,280</td>
<td>541</td>
<td>540</td>
<td>393</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>28,409</td>
<td>235</td>
<td>235</td>
<td>146</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>1,185</td>
<td>10</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>N</td>
<td>37</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>23,688</td>
<td>193</td>
<td>193</td>
<td>84</td>
</tr>
<tr>
<td>N</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>20,078</td>
<td>171</td>
<td>164</td>
<td>57</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>5,321</td>
<td>43</td>
<td>43</td>
<td>33</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>N</td>
<td>371</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>151,787</td>
<td>800</td>
<td>795</td>
<td>552</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>293,439</td>
<td>1,100</td>
<td>1,095</td>
<td>621</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>2,253</td>
<td>18</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>456</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>526,099</td>
<td>1,100</td>
<td>1,087</td>
<td>234</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>5,708,286</td>
<td>1,625</td>
<td>1,579</td>
<td>111</td>
</tr>
</tbody>
</table>

**TOTAL**

6,910,192 6,500 6,421 2,663
2009 Results (pre-adjudication)

![Graph showing precision-recall results for different topics in 2009. The x-axis represents recall, ranging from 0.0 to 1.0, and the y-axis represents precision, also ranging from 0.0 to 1.0. Each topic is represented by a different color dot: Topic 201 (2009) in blue, Topic 202 (2009) in green, Topic 203 (2009) in brown, Topic 204 (2009) in red, Topic 205 (2009) in purple, Topic 206 (2009) in yellow, and Topic 207 (2009) in light blue. The dots are scattered across the graph, indicating varying levels of precision and recall for each topic.]
2009 Results (post-adjudication)

![Graph showing the relationship between Recall and Precision for various topics in 2009. The x-axis represents Recall, ranging from 0.0 to 1.0. The y-axis represents Precision, also ranging from 0.0 to 1.0. Different colors and markers are used to distinguish the performance of topics 201 to 207 from 2009.](image-url)
2009 Results
(pre- to post- adjudication)

Recall
Precision

0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4 0.6 0.8 1.0

Topic 201 (2009)
Topic 202 (2009)
Topic 203 (2009)
Topic 204 (2009)
Topic 205 (2009)
Topic 206 (2009)
Topic 207 (2009)
First-Pass Assessor Errors
Minimizing Downside Risk

\[ \alpha = 0.05 \]
Fixed test set
Growing training set

\[ F_1 \]

\[ \tau \]

\[ \hat{F}_1 \]

\[ F_1 \]

\[ \theta \]

Annotations
Fixed test set
Growing training set

Collection = RCV1, Topic = M132, Freq = 3.33%

Stop Criterion
Success

| Desired | 95.00% |

Training documents
Minimizing Total Cost
Policies

Topic = C18
Frequency = 6.57%
E-Discovery Test Collections

- IIT CDIP (Scanned)
  - 7 million documents (no de-dupe performed)
  - Uncorrected OCR text

- EDRM Enron Version 2 (Email)
  - 455K messages (after de-dupe)
  - Extracted text

- Avocado (Email)
  - 614K messages (after de-dupe)
  - Extracted text and metadata (e.g., calendar entries)
For More Information

• FnTIR survey
  – http://ediscovery.umiacs.umd.edu

• TREC Legal Track
  – http://trec-legal.umiacs.umd.edu

• Sedona Conference Working Group 1
  – http://www.thesedonaconference.org

• DESI Workshops
  – http://www.umiacs.umd.edu/~oard/desi6/
This work has been supported in part by the National Science Foundation under grant IIS-1065250.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.