Jointly Modeling Relevance and Sensitivity for Search Among Sensitive Content

Mahmoud F. Sayed, Douglas W. Oard
I want the public to see my email. I asked State to release them. They said they will review them for release as soon as possible.

6:35 AM · Mar 5, 2015 · Twitter Web Client

7.5K Retweets 9.5K Likes

10,045 FOIA requests

~ 30k work-related emails

Exclusive: U.S. to shift 50 staff to boost office handling Clinton emails

WASHINGTON | BY ARSHAD MOHAMMED

Democratic presidential candidate Hillary Clinton speaks during a press conference at Des Moines Area Community College in Ankeny, Iowa in this August 28, 2016 file photo.

The U.S. State Department plans to move about 50 workers into temporary jobs to bolster the office sitting through Hillary Clinton's emails and grappling with a vast backlog of other requests for information to be declassified, officials said on Tuesday.
E-Discovery

1. Formulation
2. Acquisition
3. Review for Relevance
4. Review for Privilege
5. Analysis

~ 75% total cost
~ 1 month
Motivation

- **Objective** is to build “Search and Protection Engines”
  - Protect sensitive content
  - Still retrieve relevant content
  - Affordable
  - Fast

- **Motivation**
  - Review is expensive
    - Hiring law firms
  - Review is time-consuming
    - Long elapsed time between request and its response
    - Not effective access to information

Learning to Rank

Automatic Sensitivity Classification
Proposed Approaches

Prefilter

Documents → Filter → Ranker → Result

Filter → Sensitivity Classifier

Query → Result

Postfilter

Documents → Ranker → Filter → Result

Query → Sensitivity Classifier
How to evaluate such approaches?
Discounted Cumulative Gain (DCG)

**Table:**

<table>
<thead>
<tr>
<th></th>
<th>Highly Relevant</th>
<th>Somewhat Relevant</th>
<th>Not Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>+3</td>
<td>+1</td>
<td>0</td>
</tr>
<tr>
<td>Not Retrieved</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Formula:**

\[
DCG_k = \sum_{i=1}^{k} \frac{g_i}{d_i}
\]

**Example:**

- Highly Relevant: 3/1
- Somewhat Relevant: 3/2
- Not Relevant: 3/3
- 1/5

**DCG_5 = 5.7**
### Cost-Sensitive DCG (CS-DCG)

#### Table 1: Relevant vs. Not Relevant

<table>
<thead>
<tr>
<th></th>
<th>Highly Relevant</th>
<th>Somewhat Relevant</th>
<th>Not Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>+3</td>
<td>+1</td>
<td>0</td>
</tr>
<tr>
<td>Not Retrieved</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Table 2: Sensitive vs. Not Sensitive

<table>
<thead>
<tr>
<th></th>
<th>Sensitive</th>
<th>Not Sensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>-10</td>
<td>0</td>
</tr>
<tr>
<td>Not Retrieved</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Formula

\[
CS-DCG_k = \sum_{i=1}^{k} \left( \frac{g_i}{d_i} + c_i \right)
\]

#### Example

- \(CS-DCG_5 = 5.7\)
- \(CS-DCG_5 = -4.3\)
Normalized CS-DCG (nCS-DCG)

<table>
<thead>
<tr>
<th>Worst Ranking</th>
<th>Best Ranking</th>
<th>CS-DCG_worst = -19.8</th>
<th>CS-DCG_5 = -4.3</th>
<th>nCS-DCG_5 = 0.60</th>
<th>CS-DCG_5 = 5.7</th>
<th>nCS-DCG_5 = 0.71</th>
<th>CS-DCG_best = 5.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10</td>
<td>3/1</td>
<td>3/1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>3/2</td>
<td>3/2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>3/3</td>
<td>3/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Highly Relevant</th>
<th>Somewhat Relevant</th>
<th>Sensitive</th>
<th>Neither Relevant nor Sensitive</th>
</tr>
</thead>
</table>

\[ nCS-DCG\_5 = 0.60 \]
\[ nCS-DCG\_5 = 0.71 \]
Experiments
LETOR OHSUMED Test Collection

- 348,566 medical publications
  - Fields: title, abstract, Medical Subject Heading (MeSH), etc
  - 14,430 (w/rel judgements) for eval
  - 334,136 for sensitivity classifier training

- 106 queries (~150 rel judgements per query)
  - 3 levels: (2) Highly Relevant, (1) Somewhat Relevant, and (0) Not Relevant

- Simulating “sensitivity”
  - 2 MeSH labels represent sensitive content (out of 118)
    - Male Urogenital Diseases [C12]
    - Female Urogenital Diseases and Pregnancy Complications [C13]
  - 12.2% of judged documents are sensitive
Sensitivity is Topic-Dependent

![Graph showing sensitivity of topics](graph.png)

- **Easy topics**: Fewer relevant documents are sensitive.
- **Hard topics**: A larger proportion of relevant documents are sensitive.

- **Legend**:
  - Red: Relevant + Sensitive
  - Green: Relevant + Not Sensitive
nCS-DCG@10 Comparison

nCS-DCG@10

No filter | Prefilter | Postfilter

BM25 | Linear reg. | LambdaMart | AdaRank | Coor Ascent

nCS-DCG@10
Proposed Approaches

**Prefilter**
- Documents → Filter → Ranker → Result
- Sensitivity Classifier

**Postfilter**
- Documents → Ranker → Filter → Result
- Sensitivity Classifier

**Listwise LtR Optimizing nCS-DCG**

**Joint**
- Documents → Ranker → Result
- Sensitivity Classifier
- Query
nCS-DCG@10 Comparison

Listwise LtR
CS-DCG@10 Comparison

Can we reduce number of queries with negative CS-DCG scores?
Cluster-Based Replacement (CBR)

- Similar to diversity ranking
  - Retrieved documents are clustered
  - For any potentially sensitive document in the result list is replaced with a document in the same cluster but less sensitive

20 clusters using repeated bisection
## CBR Adversely Affects nCS-DCG

<table>
<thead>
<tr>
<th></th>
<th>No filter</th>
<th>Prefilter</th>
<th>Postfilter</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unclustered</td>
<td>clustered</td>
<td>unclustered</td>
<td>clustered</td>
</tr>
<tr>
<td>BM25</td>
<td>0.727</td>
<td>0.779*</td>
<td>0.800</td>
<td>0.797</td>
</tr>
<tr>
<td>Linear reg.</td>
<td>0.761</td>
<td>0.764</td>
<td>0.811*</td>
<td>0.785</td>
</tr>
<tr>
<td>LambdaMart</td>
<td>0.765</td>
<td>0.771</td>
<td>0.812*</td>
<td>0.788</td>
</tr>
<tr>
<td>AdaRank</td>
<td>0.756</td>
<td>0.779</td>
<td>0.822*</td>
<td>0.792</td>
</tr>
<tr>
<td>Coor. Ascent</td>
<td>0.762</td>
<td>0.781</td>
<td>0.816*</td>
<td>0.791</td>
</tr>
</tbody>
</table>

* Indicates two-tailed t-test with p<0.05
Conclusion

- Proposed CS-DCG and nCS-DCG to balance between relevance and sensitivity
- Joint modeling approach yields better performance than straightforward approaches
- Cluster-based replacement can reduce number of queries with negative CS-DCG scores
Next Steps

- Train a sensitivity classifier with fewer examples
- Build test collections with real sensitivities
- Experiment with tri-state classification
  - Sensitive
  - Needs human review
  - Not Sensitive
Thanks!

Mahmoud F. Sayed
mfayoub@cs.umd.edu

Data and code can be found at https://github.com/mfayoub/SASC