Indri:
Three Terabyte Tracks Later

Donald Metzler
University of Massachusetts Amherst
Center for Intelligent Information Retrieval
Indri Overview

- Off the shelf, open source retrieval engine
- Built to scale to large collections
  - Supports distributed indexing / querying
- Retrieval model
  - Combines language modeling and inference network approaches to IR
- Robust query language supports many applications
General Retrieval Strategy

<TOP>
<num> Number: 758
<title> Embryonic stem cells
...
</TOP>

#weight(
0.8 #combine(embryonic stem cells)
0.1 #combine(#1(stem cells)
   #1(embryonic stem)
   #1(embryonic stem cells))
0.1 #combine(#uw8(stem cells)
   #uw8(embryonic cells)
   #uw8(embryonic stem)
   #uw12(embryonic stem cells)))

TREC TOPIC
QUERY FORMULATOR
INDRI QUERY
INDRI
Inquiry at TREC-4

#WSUM (1.0

1.0 #WSUM (1.0

1.0 Status 1.0 of 1.0 nuclear 1.0 proliferation 1.0 treaties 1.0 violations
1.0 and 1.0 monitoring
1.0 #PHRASE( nuclear proliferation) 1.0 #PHRASE( proliferation treaties)

0.300 #3 ( long range )
0.290 #3 ( plutonium )
0.280 #3 ( international atomic energy )
0.270 #3 ( foreign relations committee )

... 
0.180 #3 ( #usa law )

...

0.030 #3 ( strategic defense initiative )
0.020 #3 ( carnegie endowment )

1.0 #WPARSUM200 (1.0

1.0 Status 1.0 of 1.0 nuclear 1.0 proliferation 1.0 treaties 1.0 violations
1.0 and 1.0 monitoring
1.0 #PHRASE( nuclear proliferation) 1.0 #PHRASE( proliferation treaties)
Proximity Formulation Evolution

- Example: “florida seminole indians”
- Bag of words
  
  \#combine( florida seminole indians )

- Statistical phrase detection
  
  \#combine( florida “seminole indians” )

- Weighted statistical phrases
  
  \#combine( \(w_T\) florida \(w_O\) “seminole indians” )

- Weighted statistical phrases
  
  \#combine( \(w_T\) florida \(w_T\) seminole \(w_T\) indians \(w_O\) “seminole indians” )
Proximity Formulation Evolution

- Weighted statistical phrases
  \[ \text{#combine}( \ w_T \text{ florida} \ w_T \text{ seminole} \ w_T \text{ indians} \ \\
  \ w_O \ #1(\text{seminole indians}) \ ) \]

- Weighted subphrases
  \[ \text{#combine}( \ w_T \text{ florida} \ w_T \text{ seminole} \ w_T \text{ indians} \ \\
  \ w_O \ #1(\text{florida seminole}) \ \\
  \ w_O \ #1(\text{seminole indians}) \ \\
  \ w_O \ #1(\text{florida seminole indians}) \ ) \]

- Weighted subsets
  \[ \text{#combine}( \ w_T \text{ florida} \ w_T \text{ seminole} \ w_T \text{ indians} \ \\
  \ w_U \ #uw8(\text{florida seminole}) \ \\
  \ w_U \ #uw8(\text{florida indians}) \ \\
  \ w_U \ #uw8(\text{seminole indians}) \ \\
  \ w_U \ #uw12(\text{florida seminole indians}) \ ) \]
Proximity Formulation Evolution

- **Dependence model**

  \[
  \texttt{combine}( \ w_T \text{ florida} \ w_T \text{ seminole} \ w_T \text{ indians} \\
  w_O \ #1(\text{florida seminole}) \\
  w_O \ #1(\text{seminole indians}) \\
  w_O \ #1(\text{florida seminole indians}) \\
  w_U \ #uw8(\text{florida seminole}) \\
  w_U \ #uw8(\text{florida indians}) \\
  w_U \ #uw8(\text{seminole indians}) \\
  w_U \ #uw12(\text{florida seminole indians}) \)
  \]

- Can be modeled in a more generally using a Markov Random Field framework
Ranking Function

MRF Scoring Function:

\[ S(D, Q) \overset{rank}{=} \lambda_T \sum_{c \in T} f_{\alpha_T}(c, D) + \lambda_O \sum_{c \in O} f_{\alpha_O}(c, D) + \lambda_U \sum_{c \in U} f_{\alpha_U}(c, D) \]

\[ \lambda_T = 0.9, \lambda_O = 0.05, \lambda_U = 0.05 \]

LM Features:

\[ f_{\mu}(c, D) = \log \frac{tf_{c,D} + \mu \frac{cf_c}{|C|}}{|D| + \mu} \]

\[ \mu_T = 1500, \mu_O = 4750, \mu_U = 4750 \]

BM25 Features:

\[ f_{k,b}(c, D) = \frac{(k + 1) tf_{c,D}}{tf_{c,D} + ((1 - b) + b \frac{|D|}{|D|_{avg}})} \]

\[ k_T = 0.9, k_O = 1.6, k_U = 1.6 \]

\[ b_T = 0.4, b_O = 0.0, b_U = 0.0 \]
## Ad Hoc Results Summary

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>BPREF</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>QL (T)</td>
<td>DM-LM (T)</td>
<td>LCE-LM (T)</td>
</tr>
<tr>
<td></td>
<td>2004 Topics (701-750)</td>
<td>0.2870</td>
<td>0.3067</td>
</tr>
<tr>
<td></td>
<td>2005 Topics (751-800)</td>
<td>0.3432</td>
<td>0.3632</td>
</tr>
<tr>
<td></td>
<td>2006 Topics (801-850)</td>
<td>0.3071</td>
<td>0.3444</td>
</tr>
<tr>
<td></td>
<td>All Topics (701-850)</td>
<td>0.3126</td>
<td>0.3383</td>
</tr>
</tbody>
</table>

**MAP**

- **QL (T)**: query likelihood
- **DM-LM (T)**: dependence model
- **LCE-LM (T)**: query expansion
- **LCE-BM25 (T)**: language modeling features
- **LCE-LM (TDN)**: BM25 features

- **T**: title only
- **TDN**: title, description, and narrative

- **BPREF**

- **P@10**

- **QL (T)**

- **DM-LM (T)**

- **LCE-LM (T)**

- **LCE-BM25 (T)**

- **LCE-LM (TDN)**
## Ad Hoc Results

<table>
<thead>
<tr>
<th>Group</th>
<th>Run</th>
<th>bpref</th>
<th>p@20</th>
<th>MAP</th>
<th>infAP</th>
<th>CPUs</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>uwaterloo-clarke</td>
<td>uwmtFadTPFB</td>
<td>0.4251</td>
<td>0.5570</td>
<td>0.3392</td>
<td>0.2999</td>
<td>1</td>
<td>964</td>
</tr>
<tr>
<td>umass.allan</td>
<td>indri06AlceB</td>
<td>0.4229</td>
<td>0.5410</td>
<td>0.3687</td>
<td>0.3157</td>
<td>1</td>
<td>38737</td>
</tr>
<tr>
<td>pekingu.van</td>
<td>TWTB06AD01</td>
<td>0.4193</td>
<td>0.5150</td>
<td>0.3737</td>
<td>0.3224</td>
<td>4</td>
<td>56160</td>
</tr>
<tr>
<td>hummingbird.tomlinson</td>
<td>humT06xle</td>
<td>0.4172</td>
<td>0.5820</td>
<td>0.3452</td>
<td>0.2947</td>
<td>1</td>
<td>36000</td>
</tr>
<tr>
<td>ibm.carmel</td>
<td>JuruTWE</td>
<td>0.4002</td>
<td>0.5670</td>
<td>0.3506</td>
<td>0.2687</td>
<td>1</td>
<td>3375</td>
</tr>
<tr>
<td>uglasgow.ounis</td>
<td>uogTB06QET2</td>
<td>0.3995</td>
<td>0.5400</td>
<td>0.3456</td>
<td>0.2861</td>
<td>1</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Ad Hoc Lessons Learned

- What Worked
  - Phrases / Term proximity
  - Document quality priors

- What Didn’t Work
  - Statistical phrases / WordNet
  - Two stage ranking

- What Could Work
  - External expansion
Named Page Finding Task

- **Document Priors**
  - Inlink count
  - PageRank

- **Document Structure**
  - Mixture of field language models

- **Term Proximity**
  - Same as used in *ad hoc* track
## Named Page Results Summary

<table>
<thead>
<tr>
<th></th>
<th>QL</th>
<th>QL-P</th>
<th>DM</th>
<th>DM-P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2005 Topics</strong>&lt;br&gt;(601-872)</td>
<td>N/A</td>
<td>0.4143</td>
<td>N/A</td>
<td>0.4405</td>
</tr>
<tr>
<td><strong>2006 Topics</strong>&lt;br&gt;(901-1081)</td>
<td>0.4634</td>
<td>0.4717</td>
<td>0.4980</td>
<td>0.5123</td>
</tr>
<tr>
<td><strong>All Topics</strong></td>
<td>N/A</td>
<td>0.4535</td>
<td>N/A</td>
<td>0.4705</td>
</tr>
</tbody>
</table>

- **QL** – mixture of field models (unigram)
- **DM** – mixture of field models (term dependence)
- **P** – Link analysis document priors
Named Page Finding: Future

- Multiple-Bernoulli language models
  - Many document fields are short (i.e. title)
  - More Boolean-like matching may be more appropriate

- Better understanding of document priors
  - Differences between WT10g and GOV2
  - Priors more or less useful?
  - Which types of priors are more useful? Why?
Conclusions

- Large collections provide interesting new modeling challenges
- New features can be used to improve effectiveness on ad hoc and named page retrieval tasks
- Many interesting modeling questions to be answered as collection sizes grow even more