Indri at Terabyte Track 2005

Donald Metzler, Trevor Strohman, Yun Zhou, W.B. Croft

University of Massachusetts, Amherst Center for Intelligent Information Retrieval
**Indri Search Engine**

- 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
Zoology 101

- The *indri* is the largest type of lemur
- When first spotted, the natives yelled “*Indri! Indri!*”
- Malagasy for "*Look! Over there!*"
Efficiency Task Setup

- **Machine configuration**
  - CPU: Intel Pentium 4 2.6 GHz x 1
  - Memory: 2GB

- **Two environments**
  - Single machine
  - Distributed across six machines

- Indexed 100% of collection
# Efficiency Task Results

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Setup</th>
<th>Index Time</th>
<th>Total Query Time (s)</th>
<th>Avg. Query Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>indri05EqIS</td>
<td>Single</td>
<td>2079</td>
<td>71700</td>
<td>1.43</td>
</tr>
<tr>
<td>indri05EqID</td>
<td>Distributed</td>
<td>327</td>
<td>24720</td>
<td>0.49</td>
</tr>
</tbody>
</table>

- **Indexing**
  - Speed: 12 GB/hr/CPU
  - Size on disk: 214 GB
  - Includes building compressed collection, document vectors

- Unofficial threaded distributed run achieves avg. query time of 0.18s

- Tradeoff between efficiency and effectiveness
Ad Hoc Task Overview

- What methods are effective for *ad hoc* retrieval on a large web collection?

**Methods**

- Dependence model
- Pseudo-relevance feedback
- Document quality prior
Dependence Model

- Term dependence shown to significantly improve effectiveness on GOV2 [Metzler SIGIR ’05]
- Model assumptions
  - Query terms occur within relevant documents
  - Query terms appear within close proximity within relevant documents
  - Subphrases within query are likely to appear in relevant documents
- Types of features
  - Single term
  - Ordered window (#1)
  - Unordered window (#uwN)
Dependence Model Example

```
#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 ) )
```
## DM vs. Query Likelihood

<table>
<thead>
<tr>
<th>Query</th>
<th>QL AvgP</th>
<th>DM AvgP</th>
<th>% vs. QL</th>
</tr>
</thead>
<tbody>
<tr>
<td>yew trees</td>
<td>0.4108</td>
<td>0.3311</td>
<td>-19.4%</td>
</tr>
<tr>
<td>animals alzheimers research</td>
<td>0.1560</td>
<td>0.1301</td>
<td>-16.6%</td>
</tr>
<tr>
<td>massachusetts textile mills</td>
<td>0.1535</td>
<td>0.1297</td>
<td>-15.5%</td>
</tr>
<tr>
<td>pennsylvania slot machine gambling</td>
<td>0.5633</td>
<td>0.4858</td>
<td>-13.8%</td>
</tr>
<tr>
<td>ephedra ma huang deaths</td>
<td>0.4653</td>
<td>0.4044</td>
<td>-13.1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>QL AvgP</th>
<th>DM AvgP</th>
<th>% vs. QL</th>
</tr>
</thead>
<tbody>
<tr>
<td>kroll associates employees</td>
<td>0.0132</td>
<td>0.0955</td>
<td>+623%</td>
</tr>
<tr>
<td>pet therapy</td>
<td>0.0650</td>
<td>0.3755</td>
<td>+478%</td>
</tr>
<tr>
<td>golden ratio</td>
<td>0.0981</td>
<td>0.2302</td>
<td>+135%</td>
</tr>
<tr>
<td>ivory billed woodpecker</td>
<td>0.3924</td>
<td>0.7051</td>
<td>+79.7%</td>
</tr>
<tr>
<td>social security means test</td>
<td>0.1232</td>
<td>0.1940</td>
<td>+57.5%</td>
</tr>
</tbody>
</table>
Why does this work?

Median = 1549

Median = 815
IR Paradigm Shifts

- **idf shift**
- **tf shift**
- **TREC Disks 1,2**
- **GOV2**
- **noise filter shift (i.e. proximity, spam)**

- Number of documents
- Average document length
Pseudo Relevance Feedback

- Query expansion based on Lavrenko’s relevance models
  - “Clipped Relevance Model” / “RM3”
- Run original query $Q_{\text{ORIG}}$ and build relevance model from top documents
- Construct expanded query $Q_{\text{RM}}$ using terms with high likelihood in relevance model
- Mix $Q_{\text{ORIG}}$ and $Q_{\text{RM}}$
- Indri query: $\texttt{#weight}(\lambda Q_{\text{ORIG}} (1 - \lambda) Q_{\text{RM}})$
Document Quality Prior

- **Hypothesis:** charts, tables, and lists are likely to be non-relevant for *ad hoc* queries
- **Idea:** create document prior that captures some notion of ‘quality’
- **Naïve Bayes classifier using two features**
  - Information-to-noise ratio
  - KL divergence with collection model
- **Mixed results**
  - Helps for simple queries
  - Hurts for more complex query formulations
## Ad Hoc Results

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Method</th>
<th>Type</th>
<th>MAP</th>
<th>BPREF</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>indri05AqI</td>
<td>QL</td>
<td>T</td>
<td>0.3252</td>
<td>0.3360</td>
<td>0.5840</td>
</tr>
<tr>
<td>indri05Adm</td>
<td>DM</td>
<td>T</td>
<td>0.3505</td>
<td>0.3610</td>
<td>0.5960</td>
</tr>
<tr>
<td>indri05AdmF</td>
<td>DM+PRF</td>
<td>T</td>
<td>0.3886</td>
<td>0.3920</td>
<td>0.6340</td>
</tr>
<tr>
<td>indri05AdmFL</td>
<td>DM+PRF</td>
<td>TDN</td>
<td>0.4041</td>
<td>0.4306</td>
<td>0.6580</td>
</tr>
</tbody>
</table>

**QL** = query likelihood  
**DM** = dependence model  
**PRF** = pseudo-relevance feedback  
**T** = title only  
**TDN** = title + description + narrative
Ad Hoc Conclusions

- Term proximity is important for task
- Effectiveness increase trends
  - Dependence model ~ 10%
  - Pseudo-relevance feedback ~ 5%
  - Combined ~ 20% (amplification)
- Mixed document quality prior results
Named Page Finding Task

- Indexed fields
  - mainbody, title, inlink, heading

- Feature-based model
  - \( S(D; Q) = w^T f(D, Q) \)
  - 26 features – PageRank + inlink count + 6 tf.idf-based features (x4 fields)

- Mixture model with document priors
  - Priors based on PageRank and inlink count
  - \( P(D \mid Q, \Theta) \propto P(D) \prod_{i} \sum_{q \in Q} P(i) P(q \mid \theta_i) \)
Inlinks Prior

![Bar graph showing the probability distribution of inlinks. The x-axis represents the logarithm of the number of inlinks, and the y-axis represents the probability density.]
PageRank Prior

![Graph showing the distribution of PageRank values. The x-axis represents PageRank values from 0 to 10, and the y-axis shows the probability density function (PDF) values ranging from 0 to 0.000007. The graph indicates a high probability density at PageRank values 8 and 9, with lower values for other PageRank values.]
## URLs with Highest PageRank

<table>
<thead>
<tr>
<th>URL</th>
<th># Inlinks</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.usgs.gov">http://www.usgs.gov</a></td>
<td>1,329,036</td>
</tr>
<tr>
<td><a href="http://www.ca.gov">http://www.ca.gov</a></td>
<td>471,819</td>
</tr>
<tr>
<td><a href="http://www.nih.gov">http://www.nih.gov</a></td>
<td>1,502,324</td>
</tr>
<tr>
<td><a href="http://www.epa.gov">http://www.epa.gov</a></td>
<td>1,386,329</td>
</tr>
<tr>
<td><a href="http://es.epa.gov/cgi-bin/ncerqamail.pl">http://es.epa.gov/cgi-bin/ncerqamail.pl</a></td>
<td>1,349,131</td>
</tr>
<tr>
<td><a href="http://es.epa.gov/ncer/rfa/">http://es.epa.gov/ncer/rfa/</a></td>
<td>1,344,630</td>
</tr>
<tr>
<td><a href="http://www.hhs.gov">http://www.hhs.gov</a></td>
<td>778,652</td>
</tr>
<tr>
<td><a href="http://www.ornl.gov">http://www.ornl.gov</a></td>
<td>639,570</td>
</tr>
<tr>
<td><a href="http://www.doi.gov">http://www.doi.gov</a></td>
<td>761,456</td>
</tr>
<tr>
<td><a href="http://www.medicare.gov">http://www.medicare.gov</a></td>
<td>186,948</td>
</tr>
</tbody>
</table>
Example Indri NP Query

```
#weight(
  0.1 #weight(
    1.0 #prior(pagerank) 0.75 #prior(inlinks) )
  1.0 #weight(
    0.9 #combine(
      #wsum( 1 stellwagen.(inlink)
             1 stellwagen.(title)
             3 stellwagen.(mainbody)
             1 stellwagen.(heading) )
      #wsum( 1 bank.(inlink)
             1 bank.(title)
             3 bank.(mainbody)
             1 bank.(heading) ) )
  0.1 #combine(
    #wsum( 1 #uw8( stellwagen bank ).(inlink)
           1 #uw8( stellwagen bank ).(title)
           3 #uw8( stellwagen bank ).(mainbody)
           1 #uw8( stellwagen bank ).(heading) ) ) )
```
## Named Page Finding Results

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Method</th>
<th>MRR</th>
<th>S@10</th>
<th>Not Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>indri05Nf</td>
<td>F</td>
<td>0.375</td>
<td>0.528</td>
<td>0.187</td>
</tr>
<tr>
<td>indri05Nmp</td>
<td>MP</td>
<td>0.414</td>
<td>0.563</td>
<td>0.175</td>
</tr>
<tr>
<td>indri05Nmpsd</td>
<td>MP+DM</td>
<td>0.441</td>
<td>0.583</td>
<td>0.171</td>
</tr>
</tbody>
</table>

F = feature-based model  
MP = mixture model + document priors  
DM = dependence model
Conclusions

- Indri both efficient and effective
- All tasks supported “out of the box” via Indri’s query language
- Term proximity important for both ad hoc and named page finding tasks
- All queries and parameter files publicly available for download
Questions?