

---

# 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

```
#combine(  $w_T$  florida  $w_T$  seminole  $w_T$  indians  
           $w_O$  #1(seminole indians) )
```

- ◆ Weighted subphrases

```
#combine(  $w_T$  florida  $w_T$  seminole  $w_T$  indians  
           $w_O$  #1(florida seminole)  
           $w_O$  #1(seminole indians)  
           $w_O$  #1(florida seminole indians) )
```

- ◆ Weighted subsets

```
#combine(  $w_T$  florida  $w_T$  seminole  $w_T$  indians  
           $w_U$  #uw8(florida seminole)  
           $w_U$  #uw8(florida indians)  
           $w_U$  #uw8(seminole indians)  
           $w_U$  #uw12(florida seminole indians) )
```

# Proximity Formulation Evolution



- ◆ Dependence model

```
#combine (  $w_T$  florida  $w_T$  seminole  $w_T$  indians
 $w_O$  #1 (florida seminole)
 $w_O$  #1 (seminole indians)
 $w_O$  #1 (florida seminole indians)
 $w_U$  #uw8 (florida seminole)
 $w_U$  #uw8 (florida indians)
 $w_U$  #uw8 (seminole indians)
 $w_U$  #uw12 (florida seminole indians) )
```

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

# Ranking Function

## MRF Scoring Function:

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

$$\lambda_T = 0.9, \lambda_O = 0.05, \lambda_U = 0.05$$

## LM Features:

$$f_{\mu}(c, D) = \log \frac{tf_{c,D} + \mu \frac{cf_c}{|C|}}{|D| + \mu} \quad \mu_T = 1500, \mu_O = 4750, \mu_U = 4750$$

## BM25 Features:

$$f_{k,b}(c, D) = \frac{(k+1)tf_{c,D}}{tf_{c,D} + \left( (1-b) + b \frac{|D|}{|D|_{avg}} \right)} \quad \begin{array}{l} k_T = 0.9, k_O = 1.6, k_U = 1.6 \\ b_T = 0.4, b_O = 0.0, b_U = 0.0 \end{array}$$

# Ad Hoc Results Summary



	MAP				
	QL (T)	DM-LM (T)	LCE-LM (T)	LCE-BM25 (T)	LCE-LM (TDN)
2004 Topics (701-750)	0.2870	0.3067	0.3326	0.3216	0.3650
2005 Topics (751-800)	0.3432	0.3632	0.4002	0.3878	0.4287
2006 Topics (801-850)	0.3071	0.3444	0.3452	0.3687	0.4252
All Topics (701-850)	0.3126	0.3383	0.35952	0.3596456	0.40655

	BPREF				
	QL (T)	DM-LM (T)	LCE-LM (T)	LCE-BM25 (T)	LCE-LM (TDN)
2004 Topics (701-750)	0.3593	0.3757	0.4071	0.4078	0.4536
2005 Topics (751-800)	0.3893	0.4064	0.4459	0.4423	0.4820
2006 Topics (801-850)	0.3662	0.3959	0.3913	0.4229	0.4747
All Topics (701-850)	0.3717	0.3928	0.41485	0.4244557	0.47022

	P@10				
	QL (T)	DM-LM (T)	LCE-LM (T)	LCE-BM25 (T)	LCE-LM (TDN)
2004 Topics (701-750)	0.5020	0.5898	0.5878	0.5776	0.6571
2005 Topics (751-800)	0.5840	0.6000	0.6300	0.6200	0.6640
2006 Topics (801-850)	0.5220	0.5620	0.5400	0.5820	0.6740
All Topics (701-850)	0.5362	0.5839	0.58591	0.5932886	0.6651

**QL** – query likelihood

**DM** – dependence model

**LCE** – query expansion

**LM** – language modeling features

**BM25** – BM25 features

**T** – title only

**TDN** – title, description, and narrative

# Ad Hoc Results



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

# 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



	<b>QL</b>	<b>QL-P</b>	<b>DM</b>	<b>DM-P</b>
2005 Topics (601-872)	N/A	0.4143	N/A	0.4405
2006 Topics (901-1081)	0.4634	0.4717	0.4980	0.5123
All Topics	N/A	0.4535	N/A	0.4705

**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