

Indri at TREC 2004: UMass Terabyte Track Overview

Don Metzler

University of Massachusetts, Amherst



Terabyte Track Summary

- GOV2 test collection
 - Collection size: 25205179 documents (426 GB)
 - Index size: 253 GB (includes compressed collection)
 - Index time: 6 hours (parallel across 6 machines) ~ 12GB/hr
 - Vocabulary size: 49,657,854
 - Total terms: 22,811,162,783
- Parsing
 - No index-time stopping
 - Porter stemmer
 - Normalization (U.S. => US, etc...)
- Topics
 - 50 .gov-related standard TREC *ad hoc* topics

UMass Runs

- indri04QL
 - query likelihood
- indri04QLRM
 - query likelihood + pseudo relevance feedback
- indri04AW
 - “adaptive window”
- indri04AWRM
 - “adaptive window” + pseudo relevance feedback
- indri04FAW
 - “adaptive window” + fields

indri04QL / indri04QLRM

- Query likelihood
 - Standard query likelihood run
 - Smoothing parameter trained on TREC 9 and 10 main web track data
 - Example:
`#combine(pearl farming)`
- Pseudo-relevance feedback
 - Estimate relevance model from top n documents in initial retrieval
 - Augment original query with these term
 - Formulation:
`#weight(0.5 #combine(QORIGINAL)
0.5 #combine(QRM))`

indri04AW / indri04AWRM

- Goal:
 - Given only a title query, automatically construct an Indri query
 - What's the best we can do without query expansion techniques?
- Can we model query term dependence using Indri proximity operators?
 - Ordered window (**#N**)
 - Unordered window (**#uwN**)

Related Work

- InQuery query formulations
 - Hand crafted formulations
 - Used part of speech tagging and placed noun phrases into **#phrase** operator
- Dependence models
 - van Risjbergen's tree dependence
 - Dependence language models
- Google



Web [Images](#) [Groups](#) ^{New!} [News](#) [Froogle](#) [more »](#)

TREC terabyte track

Search

[Advanced Search](#)
[Preferences](#)

Web

Results 1 - 10 of about 380 for TREC [terabyte track](#). (0.12 seconds)

[Terabyte TREC Homepage](#)

... this topic, with the goal of the workshop being a **TREC track** proposal for a retrieval experiment using a document collection on the order of a **terabyte** in size ...

www-nlpir.nist.gov/projects/terabyte/ - 6k - [Cached](#) - [Similar pages](#)

[Web Research Collections - Web Track](#)

... a 2005 Enterprise Search **Track** and a 2005 **Terabyte Track** are currently (Oct 2004) being considered, along with other proposals, by the TREC program committee. ...

es.cmis.csiro.au/TRECWeb/ - 4k - [Cached](#) - [Similar pages](#)

[RMIT University at TREC 2004 Terabyte Track Experiments](#)

File Format: Microsoft Powerpoint 97 - [View as HTML](#)

RMIT University at TREC 2004 **Terabyte Track** Experiments. Bodo Billerbeck, Adam Cannane, Abhijit Chattaraj, Nicholas Lester. William ...

goanna.cs.rmit.edu.au/~hugh/TREC.ppt - [Similar pages](#)

[RMIT University at TREC 2004 Terabyte Track Experiments](#)

File Format: Microsoft Powerpoint 97 - [View as HTML](#)

... Four types of retrieval topics are considered for the Heterogeneous **track** at INEX 2004. ... Indexes HTML, XML, plain text, and TREC-formatted documents. ...

goanna.cs.rmit.edu.au/~jovanp/ppt/INEX04_Het_track.ppt - [Similar pages](#)

[[More results from goanna.cs.rmit.edu.au](#)]

[Nick Craswell's HomePage](#)

... experiments. I am involved in the **TREC Terabyte Track** including the initial workshop (pdf) and formatting the **terabyte** collection itself. ...

research.microsoft.com/users/nickcr/ - 28k - Dec 16, 2004 - [Cached](#) - [Similar pages](#)

[Microsoft PowerPoint - ir16 trec2004](#)

File Format: PDF/Adobe Acrobat - [View as HTML](#)

... **Terabyte track** • TREC pushed IR research into gigabyte range – Was nearly unthinkable huge at the time – Now seems quaint • What happens at jump to ...

ciir.cs.umass.edu/cmppsci646/Slides/ir16%20trec2004.pdf - [Similar pages](#)

Assumptions

- Assumption 1
 - Query terms are likely to appear in relevant documents
- Assumption 2
 - Query term are likely to appear ordered in relevant documents
- Assumption 3
 - Query terms are likely to appear in close proximity to one another in relevant documents

Assumption 1

“Query terms are likely to appear in relevant documents”

- Proposed feature:
 - $P_{TERM}(q | \theta_D)$
(for every term q in our query)
- Indri representation:
 - \mathbf{q}
- Elements from our example:
 - **TREC**
 - **terabyte**
 - **track**

Assumption 2

“Query term are likely to appear ordered in relevant documents”

- Proposed feature:
 - $P_{ORDERED}("q_i \cdots q_{i+k}" | \theta_D)$
(for every subphrase $q_i \cdots q_{i+k}$ for $k > 1$ in our query)
- Indri representation
 - $\#1 (q_i \cdots q_{i+k})$
- Elements from our example:
 - $\#1 (\text{TREC terabyte})$
 - $\#1 (\text{terabyte track})$
 - $\#1 (\text{TREC terabyte track})$

Assumption 3

“Query terms are likely to appear in close proximity to one another in relevant documents”

- Proposed feature:
 - $P_{UNORDERED}(q_1 \cdots q_k | \theta_D)$
(for every non-singleton subset of query terms)
- Indri representation:
 - #uwN($q_1 \dots q_k$)
 - $N = 4 * k$
- Elements from our example:
 - #8(TREC terabyte)
 - #8(terabyte track)
 - #8(TREC track)
 - #12(TREC terabyte track)

Putting it all together...

- `#weight (w0 TREC`
 `w1 terabyte`
 `w2 track`
 `w3 #1 (TREC terabyte)`
 `w4 #1 (terabyte track)`
 `w5 #1 (TREC terabyte track)`
 `w6 #8 (TREC terabyte)`
 `w7 #8 (terabyte track)`
 `w8 #8 (TREC track)`
 `w9 #12 (TREC terabyte track))`
- Too many parameters!

Parameter Tying

- Assume that all weights of the same type have equal value
- Resulting query:

```
#weight( w_TERM      #combine( TREC terabyte track ) )
        w_ORDERED   #combine( #1( TREC terabyte )
                               #1( terabyte track )
                               #1( TREC terabyte track ) )
        w_UNORDERED #combine( #8( TREC terabyte )
                               #8( terabyte track )
                               #8( TREC track )
                               #12( TREC terabyte track ) ) )
```

- Now only 3 parameters!

Setting the weights

- Training data: WT10g / TREC9 + 10 web queries
- Objective function: mean average precision
- Coordinate ascent via line search
- Likely to find local maxima
 - Global if MAP is a convex function of the weights
- Requires evaluating a lot of queries
 - Not really a problem with a fast retrieval engine
- Able to use all of the training data
- Other methods to try
 - MaxEnt (maximize likelihood)
 - SVM (maximize margin)

Weights

- The following weights were found:
 - $w_{\text{TERM}} = 1.5$
 - $w_{\text{ORDERED}} = 0.1$
 - $w_{\text{UNORDERED}} = 0.3$
- Indicative of relative importance of each feature
- Also indicative of how *idf* factors over weight phrases

indri04FAW

- Combines evidence from different fields
 - Fields indexed: **anchor**, **title**, **body**, and **header** (h1, h2, h3, h4)
 - Formulation:

```
#weight ( 0.15 Q_ANCHOR
           0.25 Q_TITLE
           0.10 Q_HEADING
           0.50 Q_BODY      )
```
- Needs to be explore in more detail

Other Approaches

- Glasgow
 - Terrier
 - Divergence from randomness model
- University of Melbourne
 - Document-centric integral impacts
- CMU
 - Used Indri
- University of Amsterdam
 - Only indexed titles and anchor text
 - 20 minutes to index, 1 second to query
 - Very poor effectiveness

MAP			
fields ->	T	TD	TDN
QL	0.2565	0.2730	0.3088
QLRM	0.2529	0.2675	0.2928
AW	<i>0.2839</i>	<i>0.2988</i>	<i>0.3293</i>
AWRM	<i>0.2874</i>	<i>0.2974</i>	<i>0.3237</i>

Indri Terabyte
Track Results

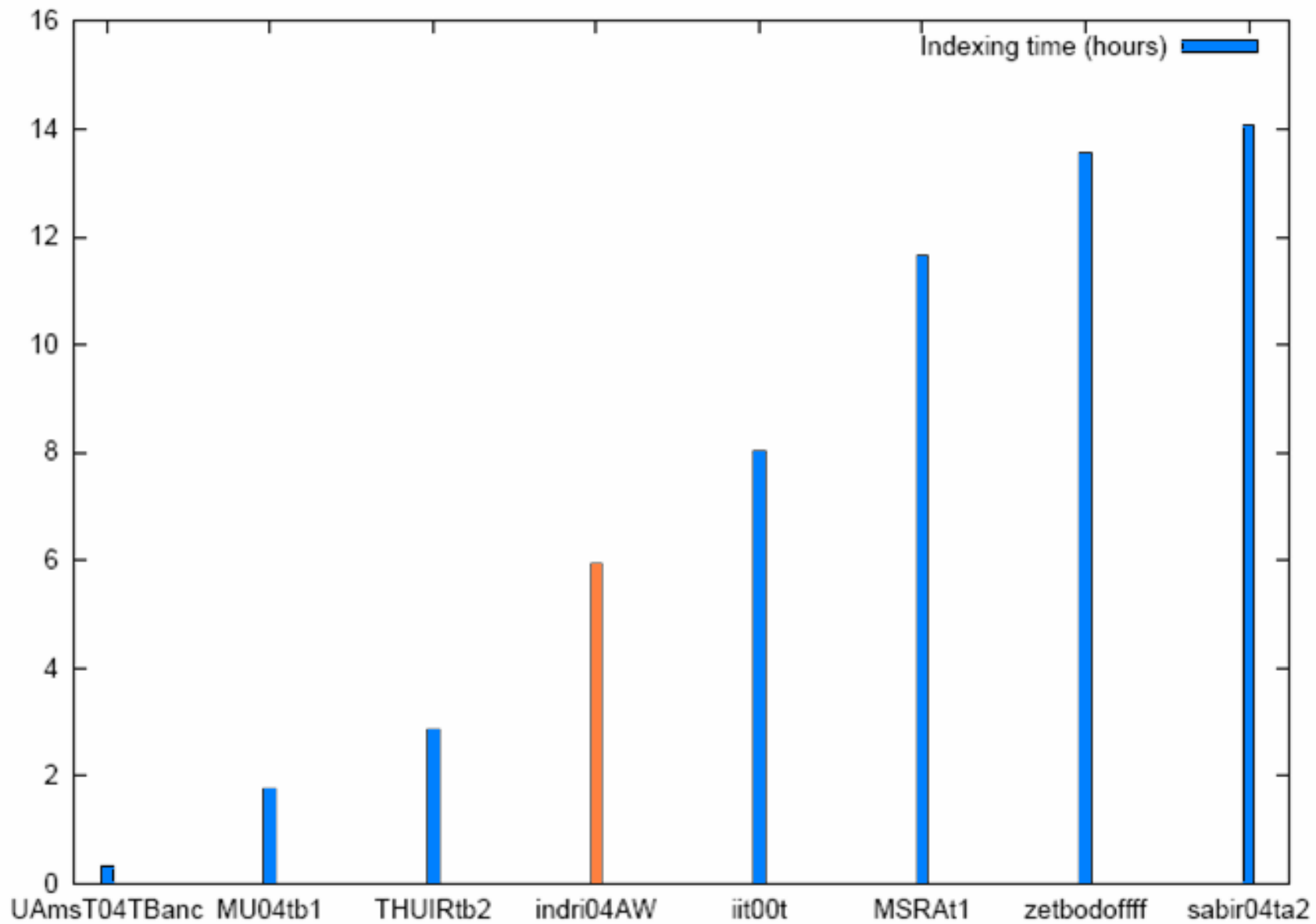
T = title

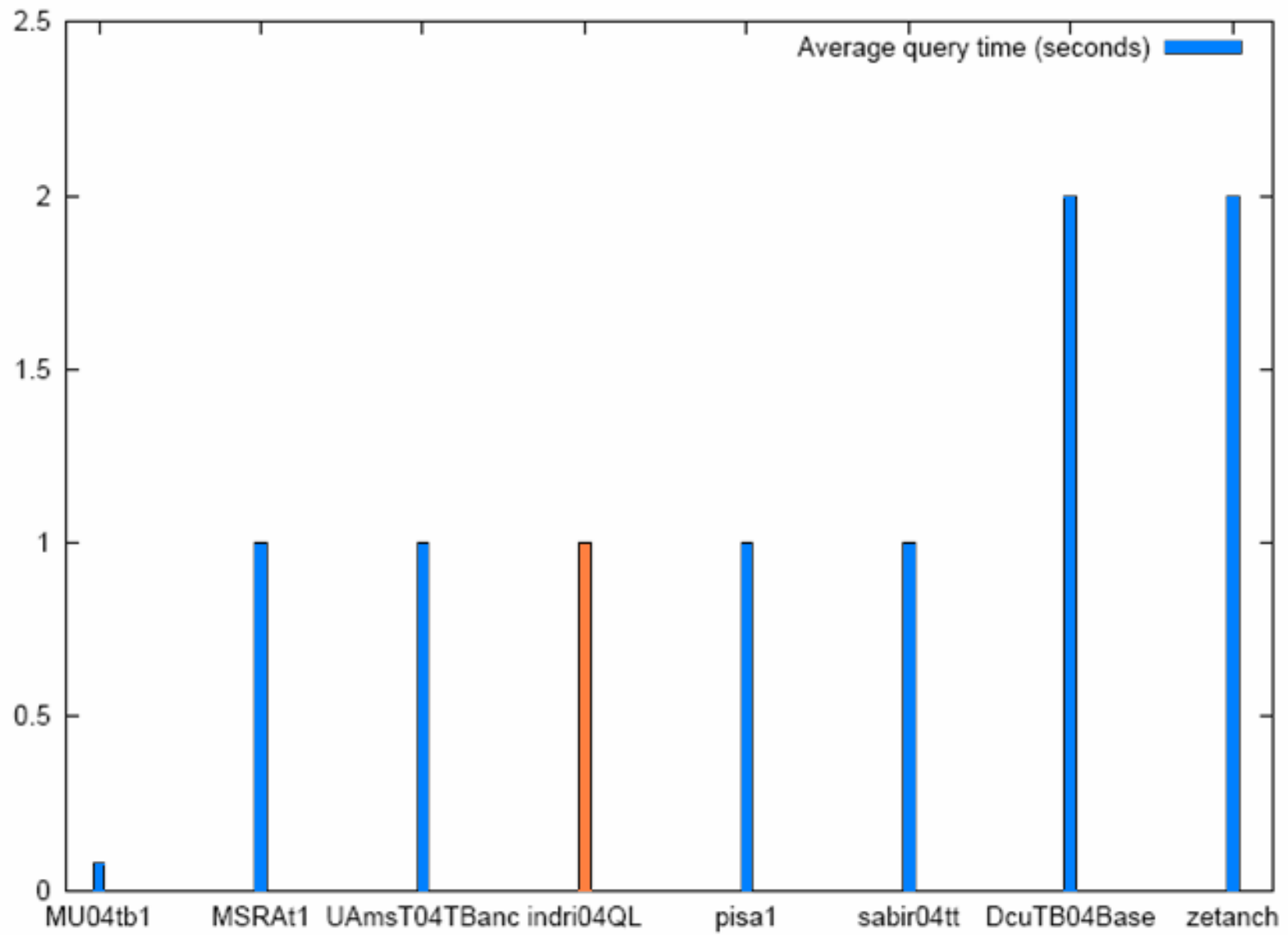
D = description

N = narrative

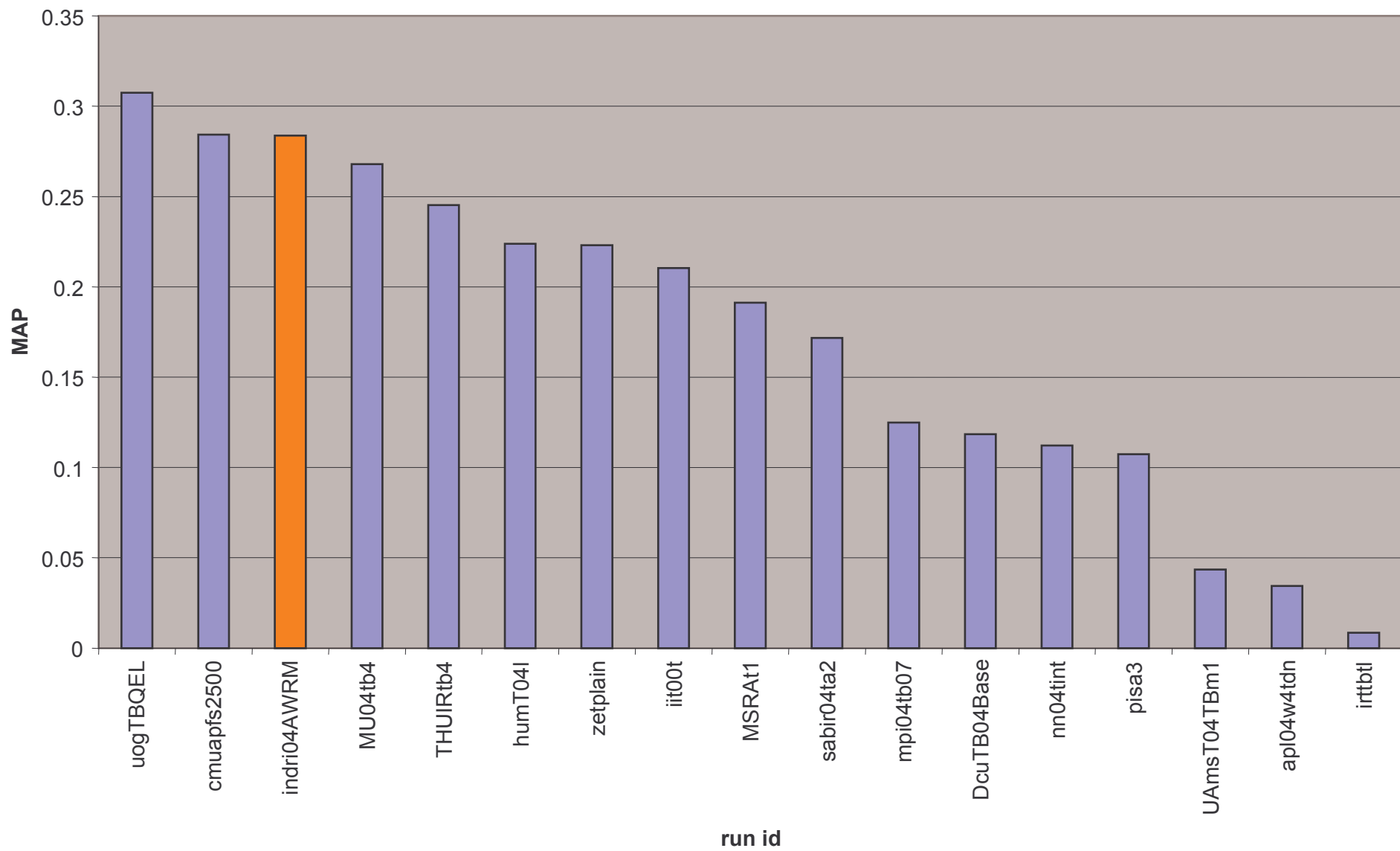
P10			
fields ->	T	TD	TDN
QL	0.4980	0.5510	0.5918
QLRM	0.4878	0.5673	0.5796
AW	<i>0.5857</i>	<i>0.6184</i>	<i>0.6306</i>
AWRM	<i>0.5653</i>	<i>0.6102</i>	<i>0.6367</i>

italicized values
denote statistical
significance over QL

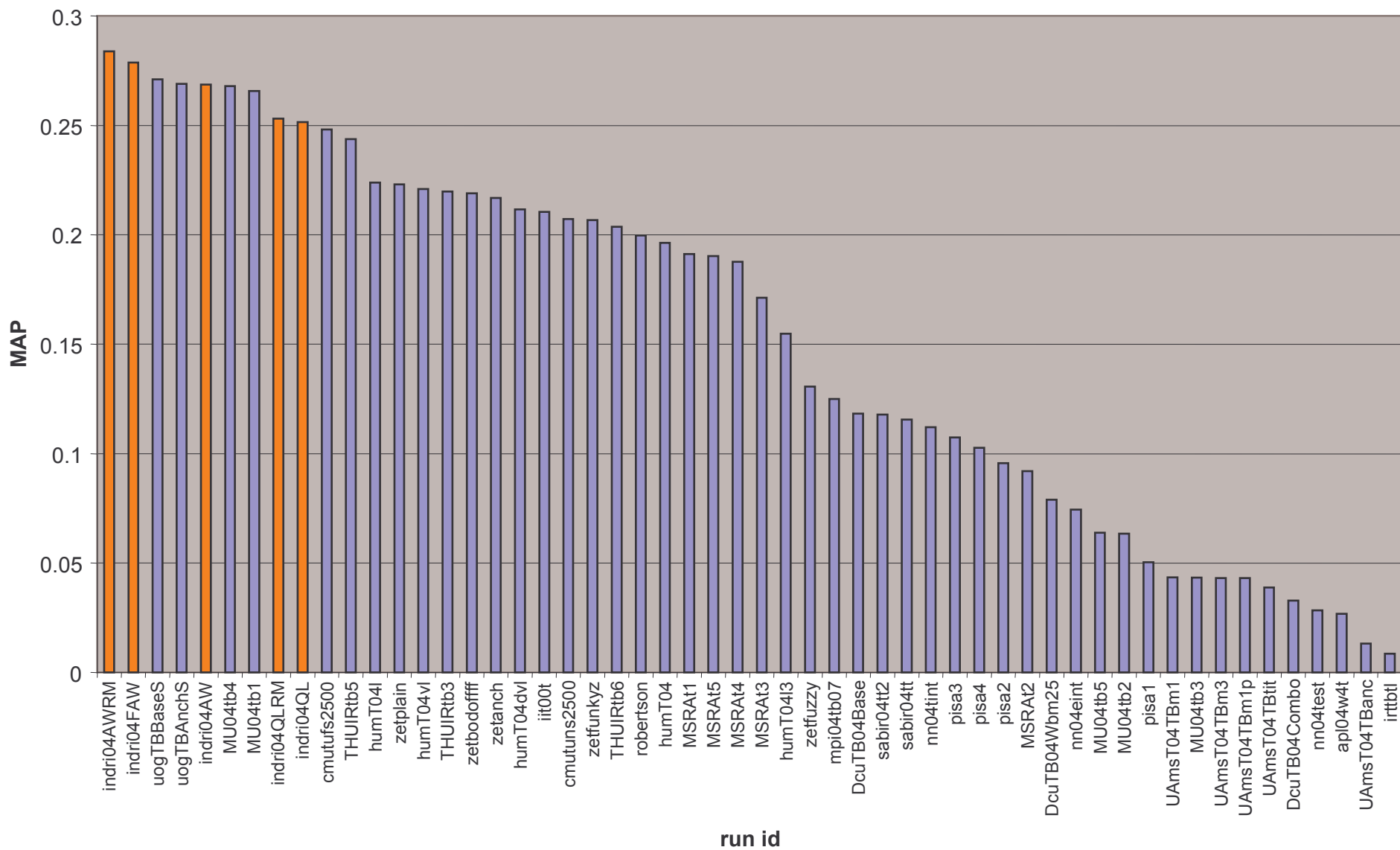




Best Run per Group



Title-Only Runs



Conclusions

- Indri was competitive both in terms of efficiency and effectiveness in the TB track arena
- “Adaptive window” approach is a promising way of formulating queries
- More work must be done on combining other forms of evidence to further improve effectiveness
 - Fields
 - Prior information
 - Link analysis