



20TH ACM CONFERENCE
ON INFORMATION AND
KNOWLEDGE MANAGEMENT

Tutorial - †AM3

Large-scale Information Retrieval Experimentation with Terrier

Rodrygo Santos, Richard McCreadie, Vassilis Plachouras

*Crowne Plaza Hotel
Glasgow, Scotland
24-28 October 2011*



www.cikm2011.org

Large-Scale Information Retrieval Experimentation with Terrier

Rodrygo L. T. Santos
rodrygo@dcs.gla.ac.uk

Richard McCreadie
richardm@dcs.gla.ac.uk

Vassilis Plachouras
vplachouras@acm.org



Presenters

- Rodrygo L. T. Santos, Univ. of Glasgow
 - Member of the Terrier Team since 2008
 - Extended Terrier to a variety of search domains
- Richard McCreadie, Univ. of Glasgow
 - Member of the Terrier Team since 2008
 - Developed Terrier's MapReduce indexing
- Dr. Vassilis Plachouras, PRESANS Inc.
 - Involved with Terrier since its inception
 - Developed much of the current core

The IR Problem

- Given an information need representation, retrieve relevant information

How can we methodically investigate new IR approaches?

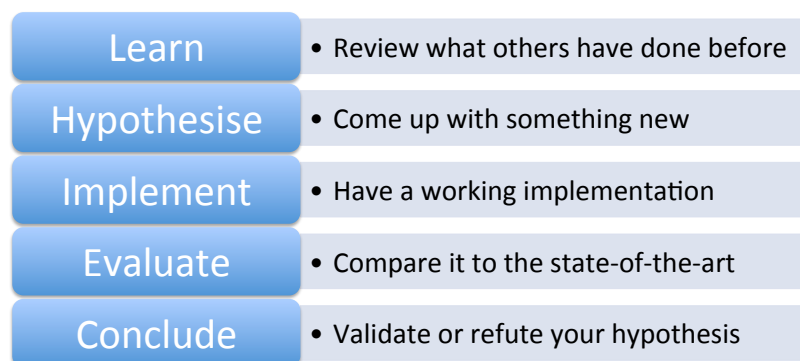
24/10/2011

© Terrier Team | University of Glasgow

3/134

IR Experimentation

- How to experiment with a new approach



24/10/2011

© Terrier Team | University of Glasgow

4/134

IR Experimentation

- Do we *really* need to do all of this?

Learn	• Review what others have done before	✓
Hypothesise	• Come up with something new	✓
Implement	• Have a working implementation	✗
Evaluate	• Compare it to the state-of-the-art	✗
Conclude	• Validate or refute your hypothesis	✓

24/10/2011

© Terrier Team | University of Glasgow

5/134

Open-Source IR Platforms

- | | |
|--|--|
| <ul style="list-style-type: none"> • Non-academic <ul style="list-style-type: none"> – Lucene / Nutch / Solr (Apache) – Minion (Oracle Labs) – Xapian (Cambridge) – Sphinx (Sphinx Inc.) | <ul style="list-style-type: none"> • Academic <ul style="list-style-type: none"> – Terrier (Glasgow) – Lemur / Indri / Galago (CMU / UMass) – Zettair (RMIT) – MG4J (Milano) |
|--|--|

24/10/2011

© Terrier Team | University of Glasgow

6/134

How Can an IR Platform Help?

- Recall the experimentation path

Implement

- Have a working implementation



- Back-end indexing provided
- Back-end retrieval provided

Evaluate

- Compare it to the state-of-the-art



- Standard baselines provided
- Evaluation tools provided

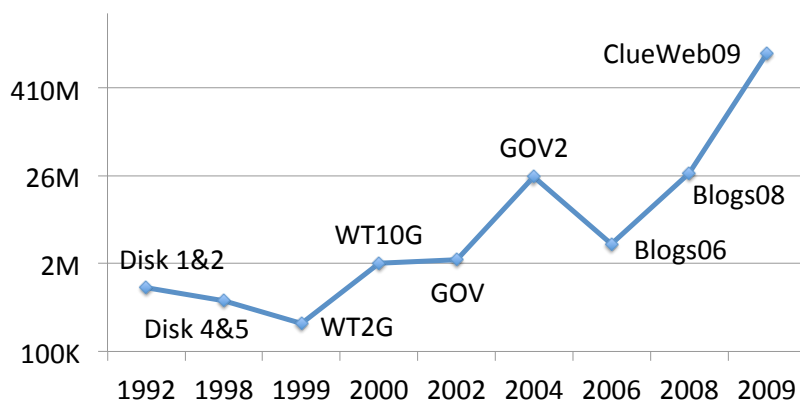
24/10/2011

© Terrier Team | University of Glasgow

7/134

IR Has Grown Big

- Increase in size



24/10/2011

© Terrier Team | University of Glasgow

8/134

IR Has Grown Big

- Increase in complexity



24/10/2011

© Terrier Team | University of Glasgow

9/134

Scaling Up IR Experimentation

- An IR platform should facilitate...
 - ... handling large corpora
 - ... tackling complex search tasks
 - ... evaluating multiple experimental scenarios

*We hope to show you that Terrier
meets all these requirements!*

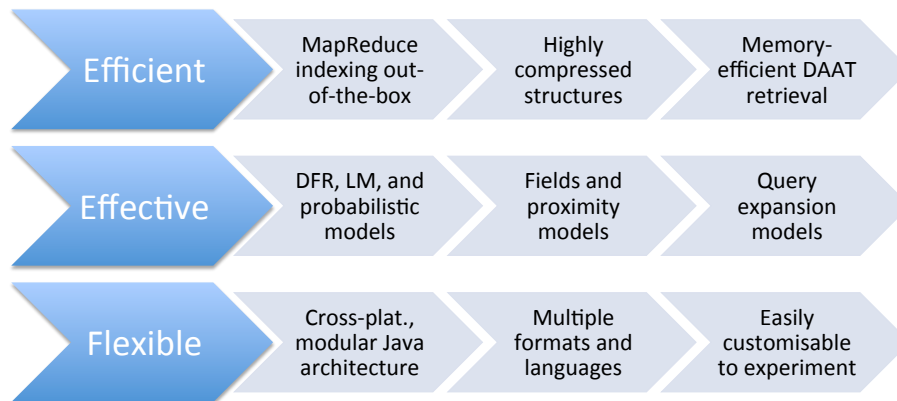
Terrier 

24/10/2011

© Terrier Team | University of Glasgow

10/134

Terrier at a Glance



24/10/2011

© Terrier Team | University of Glasgow

11/134

Learning Outcomes

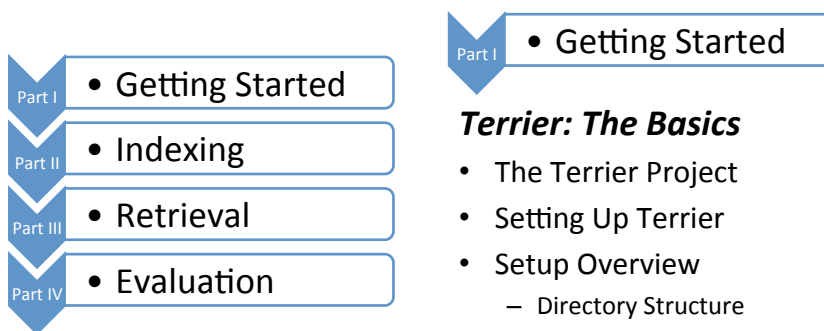
- By the end of the tutorial, you'll learn
 - The challenges of large-scale IR experimentation
 - Recent approaches to tackle size and complexity
 - How to use Terrier
 - Design and evaluate an IR experiment
 - How to extend Terrier
 - Experiment with novel research ideas
- And perhaps join us in improving Terrier! ;-)
- <http://terrier.org>

24/10/2011

© Terrier Team | University of Glasgow

12/134

Tutorial Outline



Terrier: The Basics

- The Terrier Project
- Setting Up Terrier
- Setup Overview
 - Directory Structure
 - Global Architecture

24/10/2011

© Terrier Team | University of Glasgow

13/134

Tutorial Outline



Basic Indexing

- Indexing Architecture
- Parsing Documents
- Indexing Documents
 - Single-pass Indexing
 - Position and Field Indexing
 - Metadata Indexing
- DEMO #1 (*work-along*)
 - Indexing a Collection

Extending Indexing

- Scaling Up Indexing
- The MapReduce Framework
- Terrier and MapReduce
 - Setting Up Indexing Jobs
 - Monitoring Jobs
- DEMO #2
 - Indexing with MapReduce

Coffee Break (10:30-11:00)

24/10/2011

© Terrier Team | University of Glasgow

14/134

Tutorial Outline



• Retrieval

Basic Retrieval

- Retrieval Architecture
- Retrieval Flow
- Scoring Documents
 - Several Weighting Models
 - Fields and Proximity Models
 - Query Expansion Models
- DEMO #3 (*work-along*)
 - Retrieving from an Index

Extending Retrieval

- Extending the Web Interface
- Visualising Tweet Retrieval
- DEMO #4
 - Search on the Tweets11 Corpus

24/10/2011

© Terrier Team | University of Glasgow

15/134

Tutorial Outline



• Evaluation

Batch Evaluation

- Cranfield Evaluation
 - Current Instantiations
- Batch Retrieval
- Batch Evaluation
- Scripting Evaluation
- DEMO #5 (*work-along*)
 - Evaluating in batch mode

Extending Evaluation

- Crowdsourcing
- MechanicalTurk
- Terrier and Crowdsourcing
- DEMO #6
 - Crowdsourcing Relevance Judgements

24/10/2011

© Terrier Team | University of Glasgow

16/134

Large-Scale Information Retrieval Experimentation with Terrier

PART I – GETTING STARTED

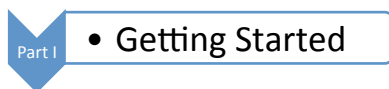
BY RODRYGO SANTOS

24/10/2011

© Terrier Team | University of Glasgow

17/134

Tutorial Outline



Terrier: The Basics

- The Terrier Project
- Setting Up Terrier
- Setup Overview
 - Directory Structure
 - Global Architecture

24/10/2011

© Terrier Team | University of Glasgow

18/134

The Terrier Project

- Research project (2001-)
 - Researchers, students, interns, and visitors
- Open-source project (2004-)
 - Latest release version 3.5 (16/06/2011)
 - Many research outcomes into the core project
- Several success cases
 - TREC Web, Robust, Terabyte, RF, MQ, Enterprise, Entity, Blog, Microblog, Crowdsourcing, Medical tracks
 - CLEF Adhoc and Web tracks
 - NTCIR Intent task

24/10/2011

© Terrier Team | University of Glasgow

19/134

Setting Up Terrier

- Single requirement
 - Java JRE 1.6.0 or higher
- Downloading Terrier

<http://terrier.org/download>

 - Unix: `terrier-3.5.tar.gz`
 - Windows: `terrier-3.5.zip`
- Ready to run!
 - No need to recompile

24/10/2011

© Terrier Team | University of Glasgow

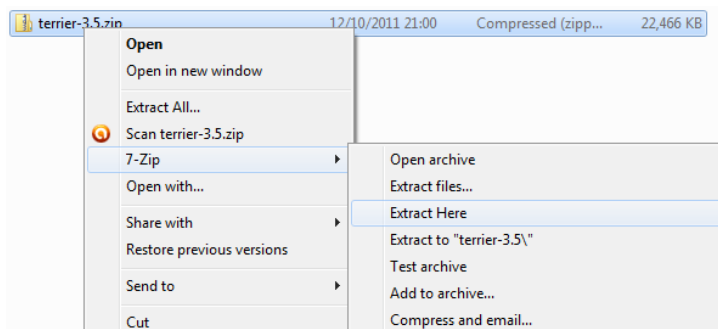
20/134

Setting Up Terrier

- Unix

```
tar zxvf terrier-3.5.tar.gz
```

- Windows



24/10/2011

© Terrier Team | University of Glasgow

21/134

Terrier Directory Structure

bin	• Utility scripts
doc	• Documentation
etc	• Configuration files
lib	• Compiled Java libraries
share	• Stopword list, tests
src	• Source code
var	• Default index and results folders

24/10/2011

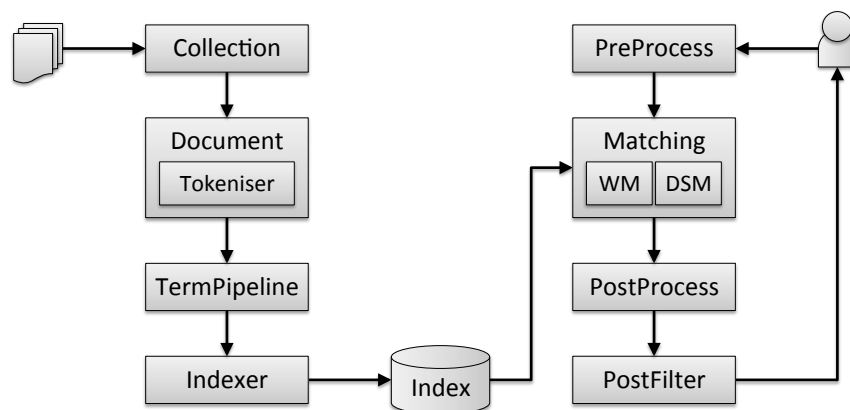
© Terrier Team | University of Glasgow

22/134

Terrier Architecture

- Indexing API

- Retrieval API



24/10/2011

© Terrier Team | University of Glasgow

23/134

Large-Scale Information Retrieval Experimentation with Terrier

PART II – INDEXING (BASICS)

BY RODRYGO SANTOS

24/10/2011

© Terrier Team | University of Glasgow

24/134

Tutorial Outline

Part II

• Indexing

Basic Indexing

- Indexing Architecture
- Parsing Documents
- Indexing Documents
 - Single-pass Indexing
 - Position and Field Indexing
 - Metadata Indexing
- DEMO #1 (*work-along*)
 - Indexing a Collection

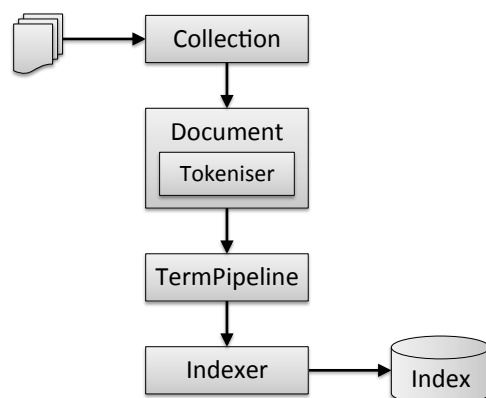
24/10/2011

© Terrier Team | University of Glasgow

25/134

Indexing Architecture

• Indexing API



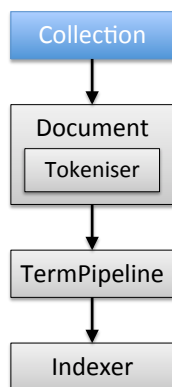
1. Collection reads documents
2. Document extracts raw text
3. Tokeniser splits extracted text
4. TermPipeline processes tokens
5. Indexer indexes processed tokens

24/10/2011

© Terrier Team | University of Glasgow

26/134

Collections



- One document per file
 - SimpleFileCollection
- Multiple documents per file
 - SimpleXMLCollection
 - TRECCollection
 - TRECWebCollection
 - WARC018Collection
 - TwitterJSONCollection

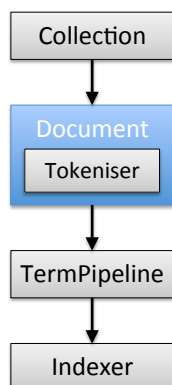
Supports ALL TREC collections!

24/10/2011

© Terrier Team | University of Glasgow

27/134

Documents



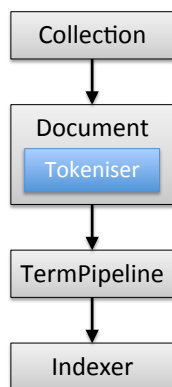
- Binary documents
 - Office: MS {Excel, Powerpoint, Word} Document
 - PDF: PDFDocument
- Text documents
 - Plain text: FileDocument
 - XML: XMLDocument
 - HTML, TREC: TaggedDocument
 - Tweets11: TwitterJSONDocument

24/10/2011

© Terrier Team | University of Glasgow

28/134

Tokenisers



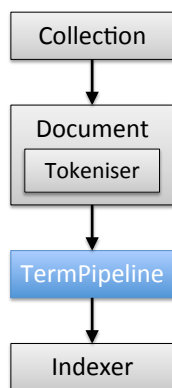
- EnglishTokeniser
 - Information Retrieval
 - information + retrieval
- ChineseTokeniser (beta)
 - 信息检索
 - 信息 + 检索
- JapaneseTokeniser (beta)
 - 情報検索
 - 情報 + 検索

24/10/2011

© Terrier Team | University of Glasgow

29/134

Term Pipelines



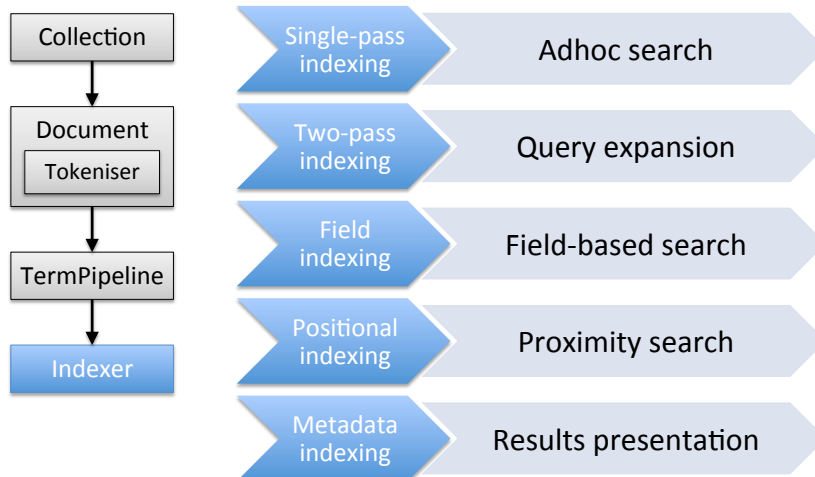
- A flexible transformation pipeline
 - Either *drop* or *transform* a token
- Stopword removal (Stopwords)
- Stemming
 - PorterStemmer
 - Full and weak versions
 - *SnowballStemmer
 - Support for 16 European languages

24/10/2011

© Terrier Team | University of Glasgow

30/134

Indexers

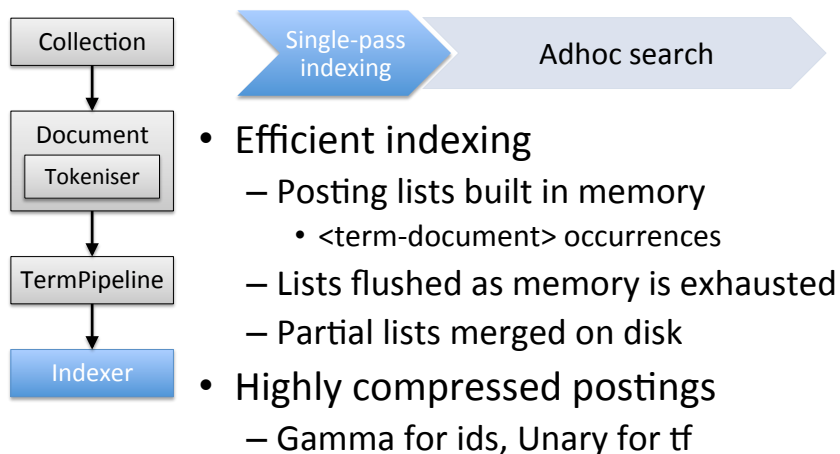


24/10/2011

© Terrier Team | University of Glasgow

31/134

Single-Pass Indexing

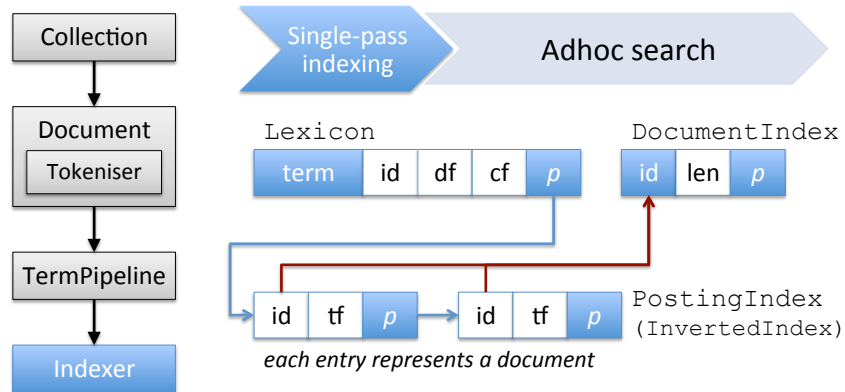


24/10/2011

© Terrier Team | University of Glasgow

32/134

Single-Pass Indexing

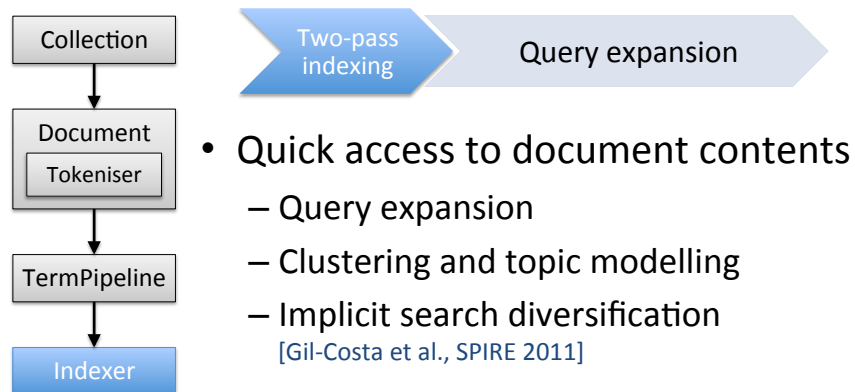


24/10/2011

© Terrier Team | University of Glasgow

33/134

Two-Pass Indexing

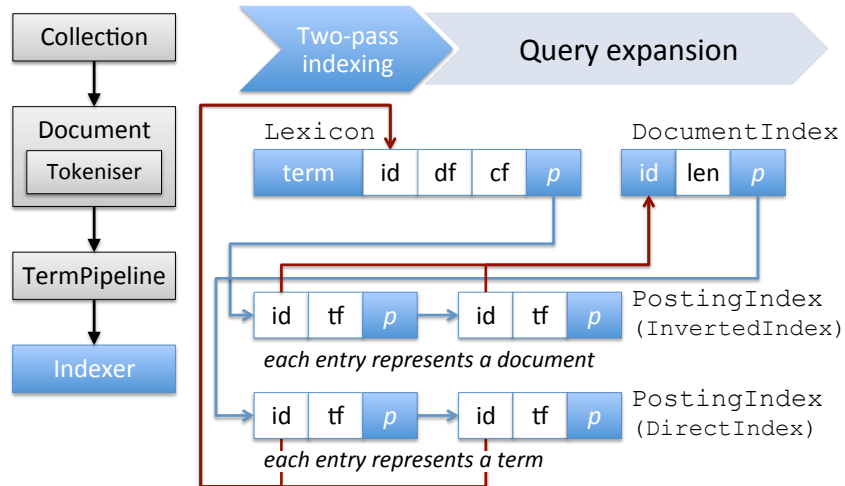


24/10/2011

© Terrier Team | University of Glasgow

34/134

Two-Pass Indexing

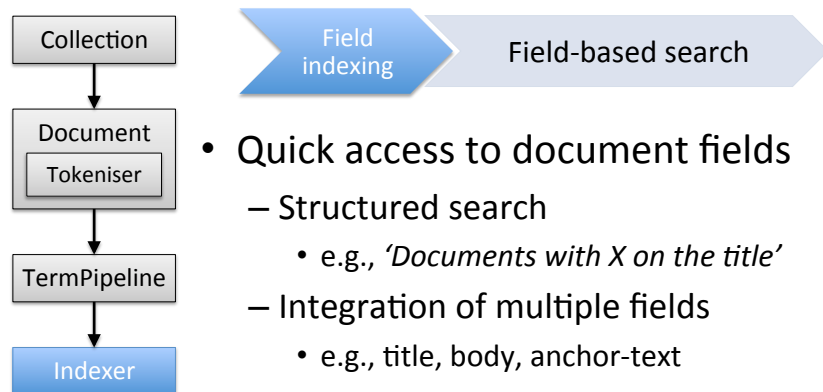


24/10/2011

© Terrier Team | University of Glasgow

35/134

Field Indexing

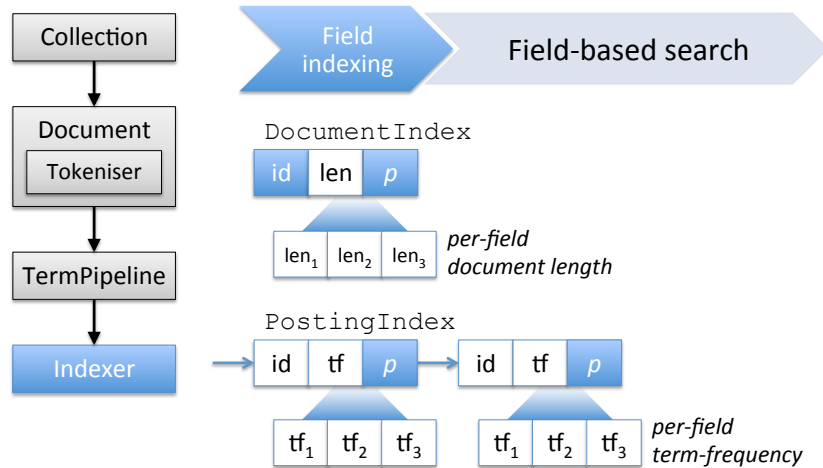


24/10/2011

© Terrier Team | University of Glasgow

36/134

Field Indexing

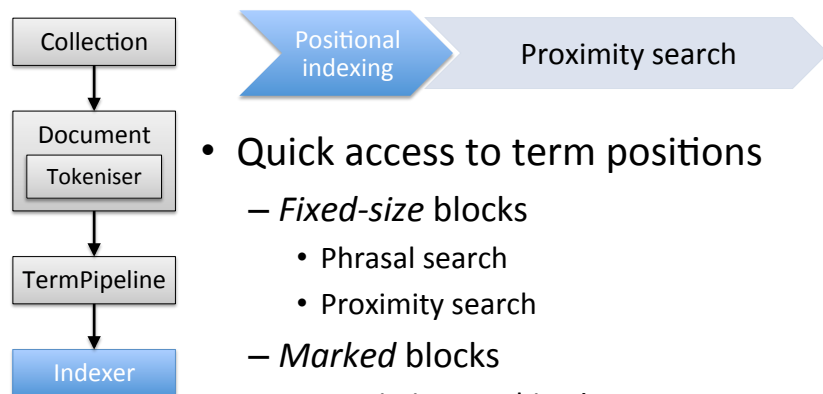


24/10/2011

© Terrier Team | University of Glasgow

37/134

Positional Indexing

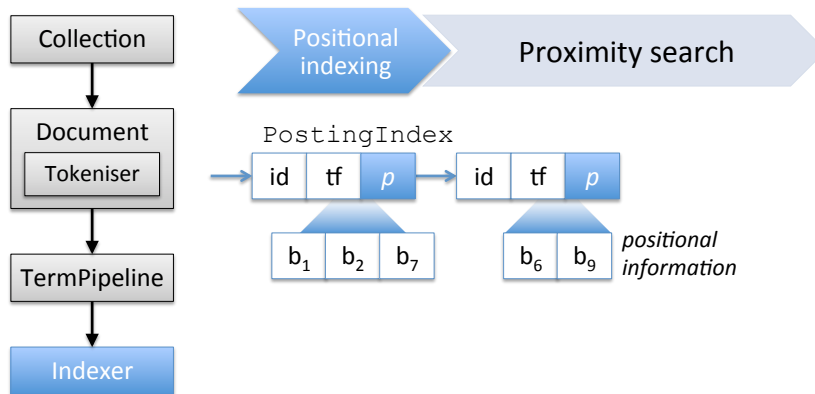


24/10/2011

© Terrier Team | University of Glasgow

38/134

Positional Indexing

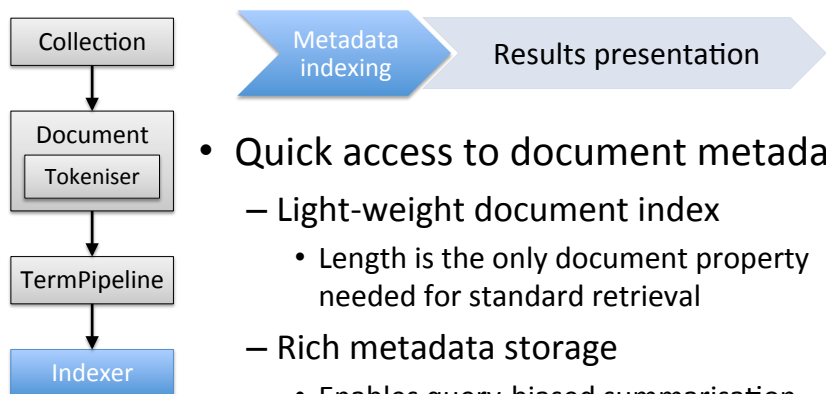


24/10/2011

© Terrier Team | University of Glasgow

39/134

Metadata Indexing

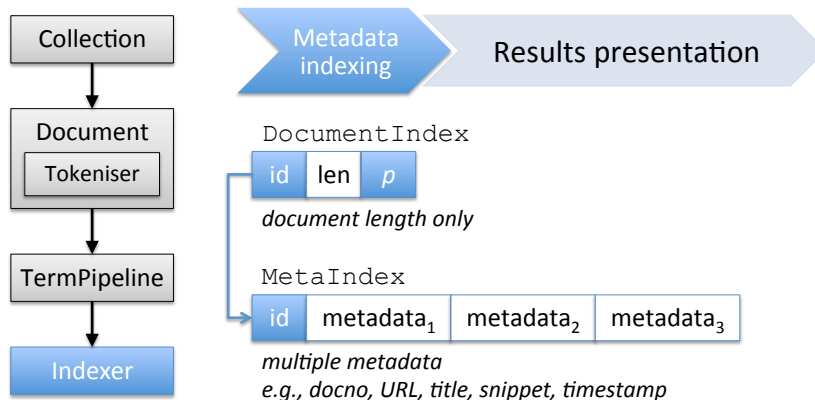


24/10/2011

© Terrier Team | University of Glasgow

40/134

Metadata Indexing



24/10/2011

© Terrier Team | University of Glasgow

41/134

DEMO #1: Indexing a Collection

```
<DOC>
<DOCNO> 0 </DOCNO>
<DOCHDR>
http://www.guardian.co.uk/music/newbandoftheday 0.0.0.0 0 text/html 0
HTTP/1.0 200 OK
Date: Fri Aug 01 15:23:59 BST 2008
Server: Apache/1.1.1
Content-type: text/html
Content-length: 0
Last-modified: Fri Aug 01 15:23:59 BST 2008
</DOCHDR>
<title> 1000 new bands of the day </title>
<summary> We celebrate our 1000th new band of the day with two free
mix tapes while new band columnist Paul Lester tells us how the
feature has ruined his life </summary>
</DOC>
```

24/10/2011

© Terrier Team | University of Glasgow

42/134

DEMO #1: Indexing a Collection

- Specify the collection to index
`bin/trec_setup.sh var/corpus`
- This will create:
 - A list of the files to index
`etc/collection.spec`
 - A default configuration properties file
`etc/terrier.properties`

24/10/2011

© Terrier Team | University of Glasgow

43/134

DEMO #1: Indexing a Collection

- Customise `etc/terrier.properties`

```
# collection specification
trec.collection.class=TRECWebCollection
# document specification
trec.document.class=TaggedDocument
# field indexing
TrecFieldTags.process=title,ELSE
# positional indexing
block.indexing=true
# metadata indexing
indexer.meta.forward.keys=docno,title,url,body
indexer.meta.forward.keylens=4,256,512,2048
TRECDocTags.propertytags=url
```

24/10/2011

© Terrier Team | University of Glasgow

44/134

DEMO #1: Indexing a Collection

- Index the specified collection

```
bin/trec_terrier.sh -i -j
```

- This will create (mainly):

```
var/index/data.document.bf  
var/index/data.inverted.bf  
var/index/data.lexicon.fsomapfile  
var/index/data.meta.zdata  
var/index/data.properties
```

24/10/2011

© Terrier Team | University of Glasgow

45/134

DEMO #1: Indexing a Collection

- Verify the index just created

```
bin/trec_terrier.sh --printstats
```

- This will show:

```
INFO - Collection statistics:  
INFO - number of indexed documents: 10000  
INFO - size of vocabulary: 34749  
INFO - number of tokens: 760213  
INFO - number of pointers: 499379
```

24/10/2011

© Terrier Team | University of Glasgow

46/134

Large-Scale Information Retrieval Experimentation with Terrier

PART II – INDEXING (MAPREDUCE)

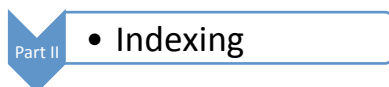
BY RICHARD MCCREADIE

24/10/2011

© Terrier Team | University of Glasgow

47/134

Tutorial Outline



Extending Indexing

- Scaling Up Indexing
- The MapReduce Framework
- Terrier and MapReduce
 - Setting Up Indexing Jobs
 - Monitoring Jobs
- DEMO #2
 - Indexing with MapReduce
- Scalability

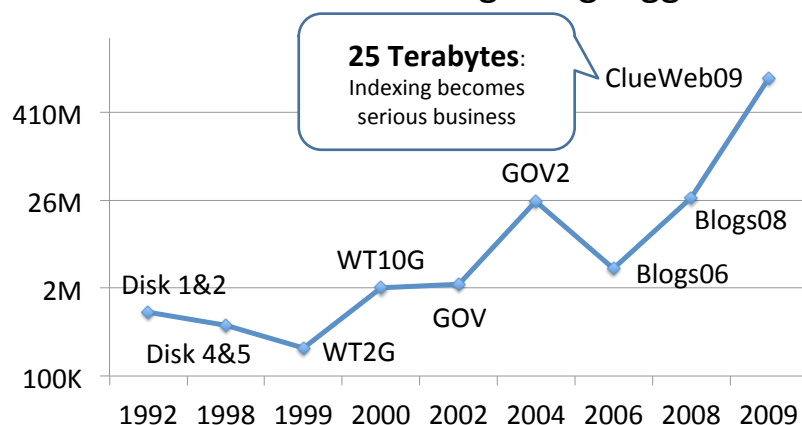
24/10/2011

© Terrier Team | University of Glasgow

48/134

Why Go Large-Scale?

- Recall: test collections are getting bigger . . .



24/10/2011

© Terrier Team | University of Glasgow

49/134

Why Go Large-Scale

- Indexing is time consuming on a single machine

Collection	Documents	Uncompressed Size (GB)	Time (minutes)
Disks 1&2	740K	2.0	8.65
Disks 4&5	500K	1.9	7.63
WT2G	240k	2.0	7.52
WT10G	1.6M	10	34.7
GOV	1.8M	18	47.1
GOV2	25M	426	1605.7

ClueWeb09?

Days or weeks

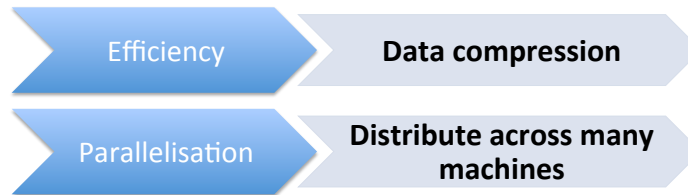
24/10/2011

© Terrier Team | University of Glasgow

50/134

Speeding Up Indexing

- Combination of:



MapReduce/  **hadoop**

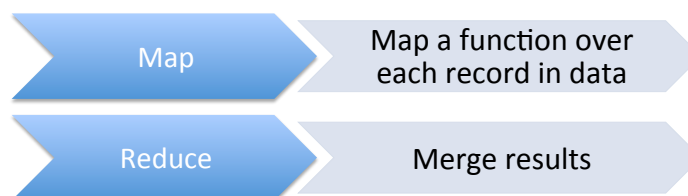
24/10/2011

© Terrier Team | University of Glasgow

51/134

What is MapReduce?

- Framework for distributing computation over multiple machines
 - Idea: Many tasks involve doing a simple map operation over each record in a large dataset



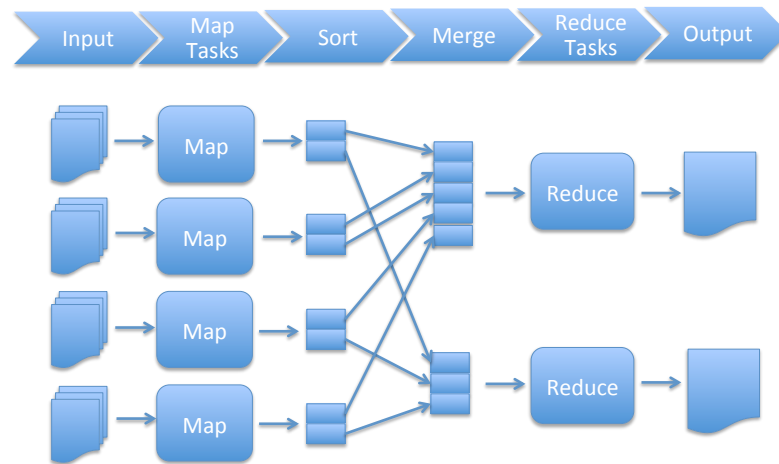
[Dean & Ghemawat, OSDI 2004]

24/10/2011

© Terrier Team | University of Glasgow

52/134

How Does it Work?



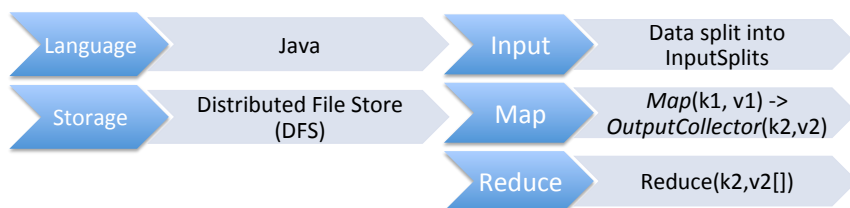
24/10/2011

© Terrier Team | University of Glasgow

53/134



- An open-source Java implementation of MapReduce by **YAHOO!**



[The Apache Hadoop Project
<http://hadoop.apache.org>]

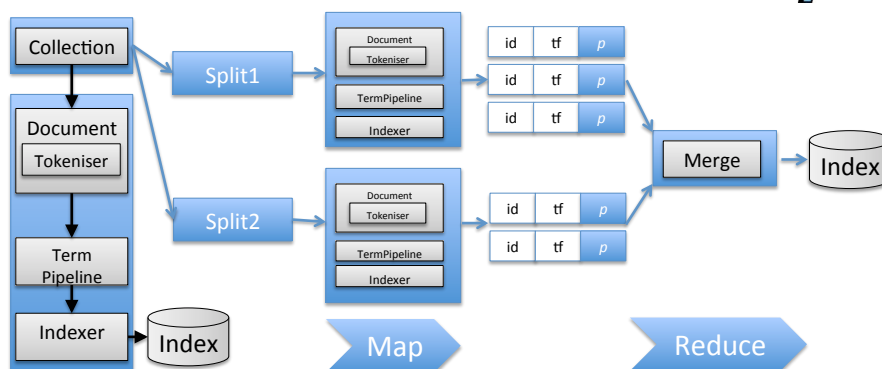
24/10/2011

© Terrier Team | University of Glasgow

54/134

MapReduce Indexing using Terrier

- The open source release of Terrier contains a distributed indexer built upon  **hadoop**

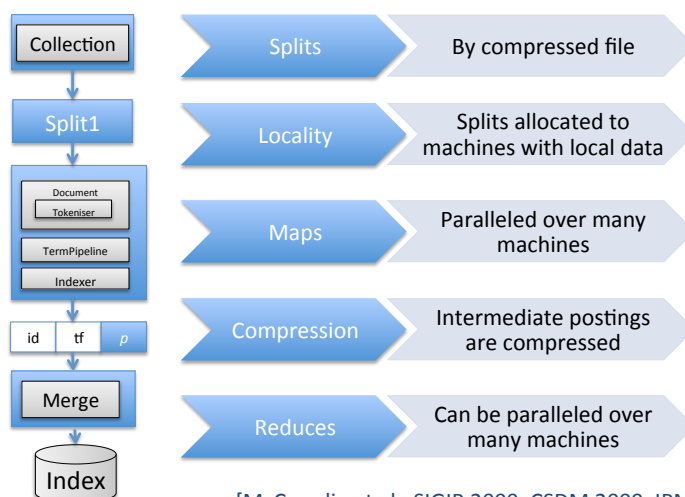


24/10/2011

© Terrier Team | University of Glasgow

55/134

Efficiency & Parallelism



[McCreadie et al., SIGIR 2009, CSDM 2009, IPM 2011]

24/10/2011

© Terrier Team | University of Glasgow

56/134

What Do I Need to Use It?

1. A cluster of machines, the more the better
2. Either:
 - The cluster dedicated to  **hadoop** (v0.20.x)
 - HadoopOnDemand (HOD) installed for on-the-fly machine allocation via Torque PBS
3. Hadoop distributed file system (HDFS) installed
4. An instance of Terrier  configured to use your Hadoop cluster

24/10/2011

© Terrier Team | University of Glasgow

57/134

How Do I Use It?

- Set up your Hadoop cluster:
 - [[Hadoop Documentation](http://hadoop.apache.org/core/docs/current): <http://hadoop.apache.org/core/docs/current>]
- Move the collection to be indexed to the HDFS
- Configure Terrier:
 - Add the location of your `$HADOOP_HOME/conf` folder to the CLASSPATH
 - Add to your `terrier.properties` file


```
terrier.plugins=org.terrier.utility.io.HadoopPlugin
```

 - [[Configuring Terrier for Hadoop](http://terrier.org/docs/v3.5/hadoop_configuration.html)
http://terrier.org/docs/v3.5/hadoop_configuration.html]
- Run `bin/trec_terrier.sh -i -H`

24/10/2011

© Terrier Team | University of Glasgow

58/134

DEMO #2: MapReduce Indexing

- <http://MASTER:PORT/jobtracker.jsp>

trnode25 Hadoop Map/Reduce Administration

State: RUNNING

Started: Wed Oct 12 13:57:33 BST 2011

Version: 0.20.2+228, rfc3233ec0769b11a99add328261295aaf4d1ad

Compiled: Sun Mar 21 23:15:39 PDT 2010 by chad

Identifier: 201110121357

Cluster Summary (Heap Size is 53.12 MB/888.94 MB)

Maps	Reduces	Total Submissions	Nodes	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node	Blacklisted Nodes
21	21	5	7	21	21	6.00	0

Scheduling Information

Queue Name	Scheduling Information
default	N/A

Filter (Jobid, Priority, User, Name)

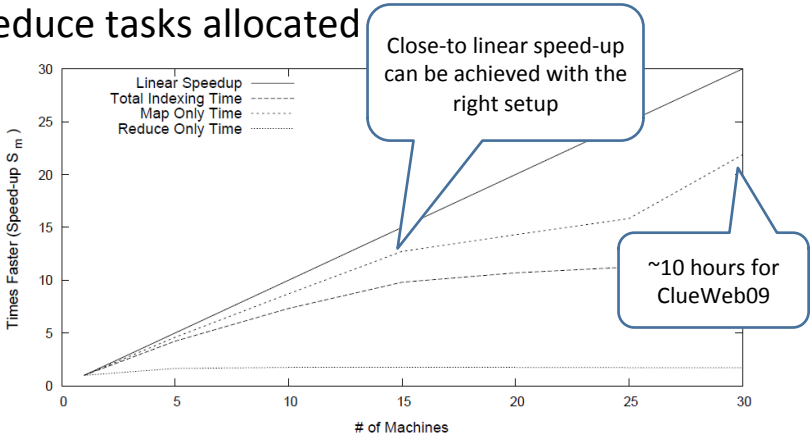
Example: "user:smith 3200" will filter by "smith" only in the user field and "3200" in all fields.

Running Jobs

Jobid	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed
job_201110121357_0006	NORMAL	richardm	terrierindexing	11.63%	498	42	2.27%	26	0

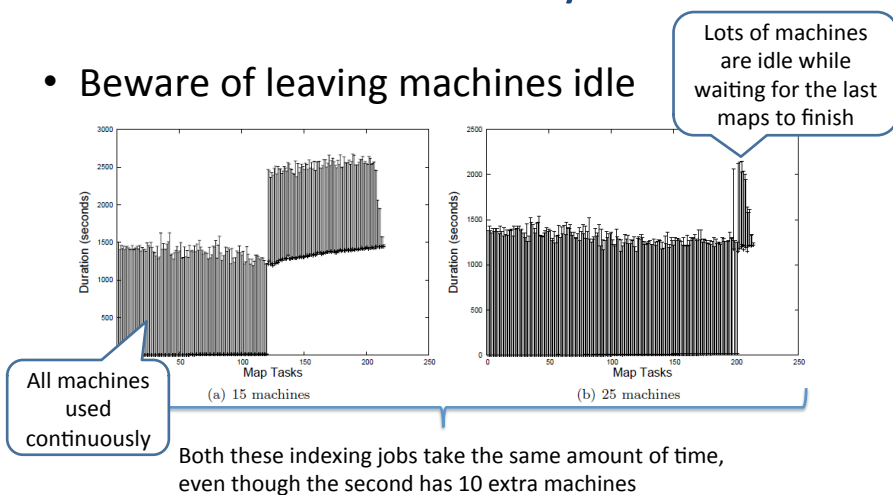
Scalability

- Dependant upon the number of Map and Reduce tasks allocated



Scalability

- Beware of leaving machines idle



24/10/2011

© Terrier Team | University of Glasgow

61/134

Large-Scale Information Retrieval Experimentation with Terrier

IT'S COFFEE TIME! =D
(resuming at 11:00)

24/10/2011

© Terrier Team | University of Glasgow

62/134

Large-Scale Information Retrieval Experimentation with Terrier

PART III – RETRIEVAL (BASICS)

BY VASSILIS PLACHOURAS

24/10/2011

© Terrier Team | University of Glasgow

63/134

Tutorial Outline



Basic Retrieval

- Retrieval Architecture
- Retrieval Flow
- Scoring Documents
 - Several Weighting Models
 - Fields and Proximity Models
 - Query Expansion Models
- DEMO #3 (*work-along*)
 - Retrieving from an Index

24/10/2011

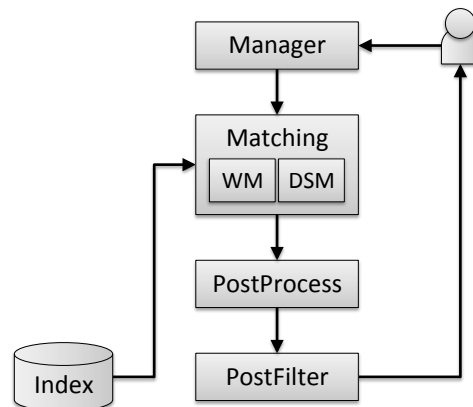
© Terrier Team | University of Glasgow

64/134

Retrieval Architecture

1. Manager parses user query
2. Matching retrieves documents
3. WMs and DSMs score documents
4. PostProcess updates result set as a whole
5. PostFilter updates individual retrieved documents

• Retrieval API

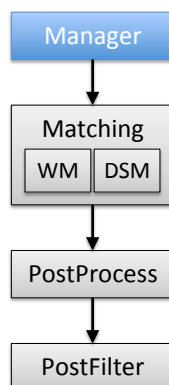


24/10/2011

© Terrier Team | University of Glasgow

65/134

Querying Manager



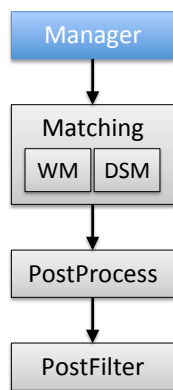
- Parses queries
- Applies TermPipeline (same as in indexing)
 - removing stopwords
 - applying stemming

24/10/2011

© Terrier Team | University of Glasgow

66/134

Querying Manager



- Terrier query language

`electric vehicle`
 retrieve documents with `electric` or `vehicle`

`toyota^2 ford`
 retrieve documents with `toyota` or `ford` where
`toyota` is weighted twice as much as `ford`

`+toyota -ford`
 retrieve documents with `toyota` but not `ford`

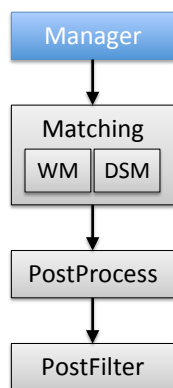
`+(toyota ford)`
 retrieve documents with both `toyota` and `ford`

24/10/2011

© Terrier Team | University of Glasgow

67/134

Querying Manager



- Terrier query language

`"electric vehicle"`
 retrieve documents with `"electric vehicle"` as a
 phrase (requires indexing with positional information)

`"electric vehicle"~5`
 retrieve documents where `electric` and `vehicle`
 appear within the given distance

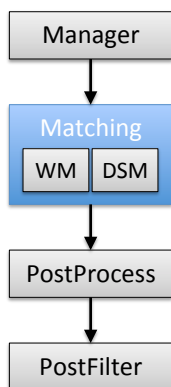
`title:ford`
 retrieve documents where `ford` appears in the title

24/10/2011

© Terrier Team | University of Glasgow

68/134

Matching



- Configurable matching strategies

- Term-at-a-time (TAAT)

- Process query terms one by one
 - Accumulate partial scores

- Document-at-a-time (DAAT)

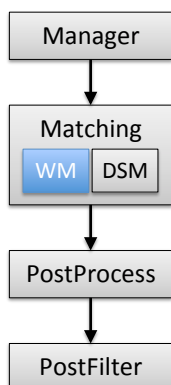
- Process documents one by one
 - Parallel scanning of posting lists
- Requires less memory than TAAT
- Enables dynamic pruning (beta)
 - [Macdonald et al., TOIS 2011]

24/10/2011

© Terrier Team | University of Glasgow

69/134

Weighting Models



- Divergence From Randomness (DFR)
 - [Amati, PhD Thesis 2003]

$$w_{d,q} = \sum_{t \in q \cap d} (-\log P_1) \cdot (1 - P_2)$$

- 3 components

- **Randomness model** computes P_1 , (e.g. Binomial distribution)
- **Risk/Information Gain Model** computes P_2 (e.g. Laplace after-effect model)
- **Term Frequency Normalisation** (e.g. Normalisation 2)

$$tfn_{d,t} = tf_{d,t} \cdot \log_2 \left(1 + c \cdot \frac{l}{l_d} \right)$$

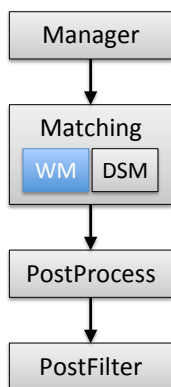
Generates $8 \times 4 \times 11 = 352$ DFR models!

24/10/2011

© Terrier Team | University of Glasgow

70/134

Weighting Models



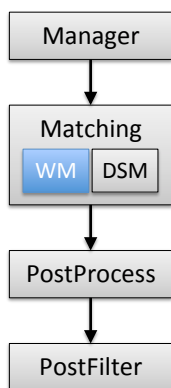
- Standard DFR models [Amati, PhD Thesis 2003]
 - BB2, IFB2, I(ne)C2, I(ne)B2, InL2, PL2
- Parameter-free DFR models
 - DLH, DLH13, XSqrA_M [Amati et al., ECIR 2006]
 - DPH [Amati et al., TREC 2007]
- Log-logistic DFR models
 - LGD [Clinchant & Gaussier, SIGIR 2010]
- Divergence From Independence (DFI)
 - DFI0 [Dincer et al., TREC 2009]
- Several classical models
 - BM25, TF-IDF, LM (Dirichlet, Hiemstra)

24/10/2011

© Terrier Team | University of Glasgow

71/134

Field-based Weighting Models



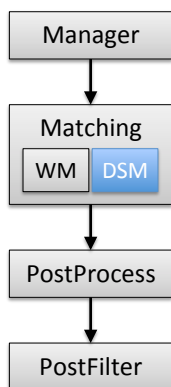
- Score terms occurrences according to
 - Per-field frequency (e.g. title, body, anchor text)
 - Per-field length normalisation
- Several field-based models
 - PL2F: per-field normalization model 2F [Macdonald et al., CLEF 2005]
 - BM25F: extension of BM25 [Zaragoza et al., TREC 2004]
 - ML2, MLD2: multinomial DFR models [Plachouras & Ounis, ECIR 2007]
- Parameters depend on specific models
 - Weight of i -th field: w_i
 - Length normalization for PL2F, BM25F: c_i

24/10/2011

© Terrier Team | University of Glasgow

72/134

Document Score Modifiers



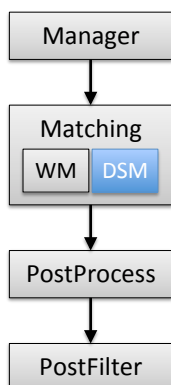
- Modify the scores assigned to a retrieved document
 - Statically configured DSMs
 - Applied for all queries
`BooleanScoreModifier: match`
documents containing all terms
 - Query-specific DSMs
 - Enabled when required for a query
`PhraseScoreModifier: applied for`
phrase queries ("electric vehicle")

24/10/2011

© Terrier Team | University of Glasgow

73/134

DSMs for Proximity Search



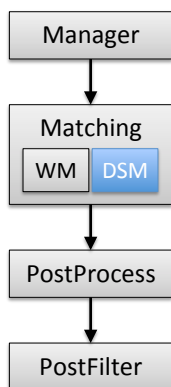
- Compute a term proximity score
 - Scores pairs of terms co-occurring in a window of text
- Available implementations
 - DFR term dependence model
[\[Peng et al., SIGIR 2007\]](#)
 - Markov random fields model
[\[Metzler & Croft, SIGIR 2005\]](#)

24/10/2011

© Terrier Team | University of Glasgow

74/134

DSMs for Prior Integration



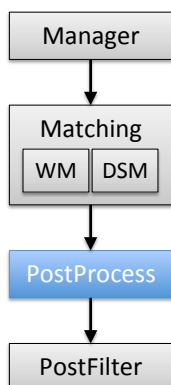
- Integrate an estimate of prior relevance
 - Query-independent document score
e.g., PageRank, readability score
- SimpleStaticScoreModifier
 - Loads external file of values
 - Adds query-independent score to the query-dependent relevance score
- Highly configurable
 - Path of input file
 - Input file format
 - Combination weight

24/10/2011

© Terrier Team | University of Glasgow

75/134

Post Processes



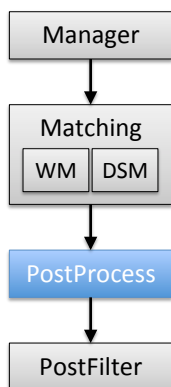
- Alter retrieved result set as a whole
 - Relevance Feedback
 - Reformulate query with evidence from relevant and non-relevant documents
 - Interactive, or blind (pseudo-relevance feedback)
 - Terrier Query Expansion
 - FeedbackSelector
 - Select top-*k* documents
 - Select relevant docs from file
 - Apply a query expansion model
 - Process the re-formulated query

24/10/2011

© Terrier Team | University of Glasgow

76/134

Query Expansion Models



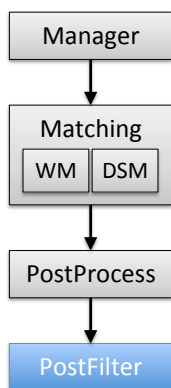
- Terms weighted by the divergence between their distributions
 - In the (pseudo-)feedback set
 - In the collection (a **random distribution**)
- Implemented models [Amati, PhD Thesis 2003]
 - Kullback-Leibler divergence: KL, KLComplete, KLCorrect
 - Bose-Einstein statistics: Bo1, Bo2
 - Chi-Square divergence: CS, CSCorrect

24/10/2011

© Terrier Team | University of Glasgow

77/134

Post Filters



- Alter or drop individual documents from the retrieved result set
 - Decorate
 - Decorate results with metadata e.g., URL, title, summary
 - SiteFilter
 - Remove results outside a give domain

24/10/2011

© Terrier Team | University of Glasgow

78/134

Interactive Retrieval with Terrier

- Run the script `bin/interactive_terrier.sh`

```
$ bin/interactive_terrier.sh
Setting TERRIER_HOME to /tmp/terrier-3.5
INFO - Structure meta reading lookup file into memory
INFO - Structure meta reading reverse map for key docno directly from
disk
INFO - Structure meta loading data file into memory
INFO - time to initialise index : 0.253
Please enter your query: Divergence
      Displaying 1-25 results
0 951 930 5.305100154627915
1 525 504 4.649211447850058
2 1129 1108 4.518302443765654
....
23 507 486 0.8871309617564661
24 48 30 -0.42021498440995214
```

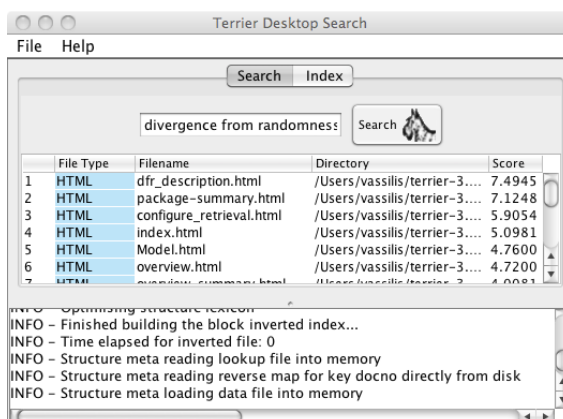
24/10/2011

© Terrier Team | University of Glasgow

79/134

Interactive Desktop Search

- Run the script `bin/desktop_terrier.sh`
- In "Index" tab, select folders to index
- In "Search" tab, submit queries



24/10/2011

© Terrier Team | University of Glasgow

80/134

DEMO #3: Retrieving from an Index

- Set up retrieval properties

```
trec.collection.class=TRECWebCollection
TaggedDocument.abstracts=title,body
TaggedDocument.abstracts.tags=title,ELSE
TaggedDocument.abstracts.lengths=120,500
indexer.meta.forward.keys=docno,title,body,url
indexer.meta.forward.keylens=32,120,500,140

querying.postfilters.order=Decorate
querying.postfilters.controls=decorate:Decorate
querying.default.controls=start:0,end:9,
    decorate:on,summaries:body,emphasis:title,body
querying.allowed.controls=scope,qe,qemodel,start,end
```

24/10/2011

© Terrier Team | University of Glasgow

81/134

DEMO #3: Retrieving from an Index

- Launch Web Terrier

```
bin/http_terrier.sh 8080 src/webapps/cikmnews
```

- Query Web interface

- Open browser and go to

```
http://localhost:8080/
```

- Submit queries

- electric vehicle
- electric -vehicle
- "electric vehicle"

24/10/2011

© Terrier Team | University of Glasgow

82/134

Large-Scale Information Retrieval Experimentation with Terrier

PART III – RETRIEVAL (INTERFACE)

BY RICHARD MCCREADIE

24/10/2011

© Terrier Team | University of Glasgow

83/134

Tutorial Outline



- Retrieval

Extending Retrieval

- Extending the Web Interface
- Visualising Tweet Retrieval
- DEMO #4
 - Search on the Tweets11 Corpus

24/10/2011

© Terrier Team | University of Glasgow

84/134

Web Interface

- Terrier comes with a search front end for:
 - Serving up Web-pages search engine style
 - Allowing researchers an easy means to visualise results from new ranking approaches

Terrier 

Estonia economy BM25 Search

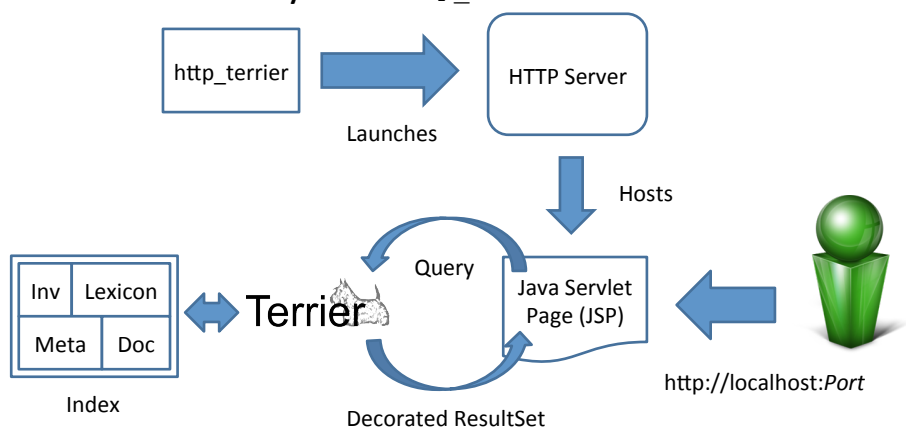
Results for Estonia economy, displaying 1-11 of 18047

Rank	Document	Score
1	http://server.soros.org/80/estonia/estocoun.html Open Estonia Foundation - About the Country server.soros.org/80/estonia/estocoun.html - WT09-B18-220 - 853062034000	11.148
2	Significance of the Estonian-Russian Interest Rate Differential There are few better proxies for measuring economic and political risk in the former Soviet Union than the interest rate differential between Estonia and Russia www.kemper.com/80/lite/curious/global_economy/weekly_report/estonia1.html - WT01-B13-137 - 852570874000	11.046

24/10/2011 © Terrier Team | University of Glasgow 85/134

How Does It Work?

- Launched by: `bin/http_terrier.sh Port Interface`

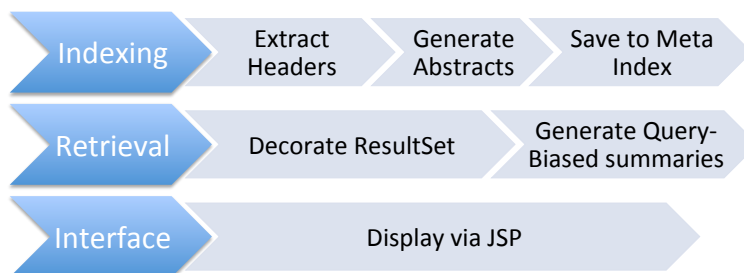


24/10/2011

© Terrier Team | University of Glasgow

86/134

Web Interface Setup



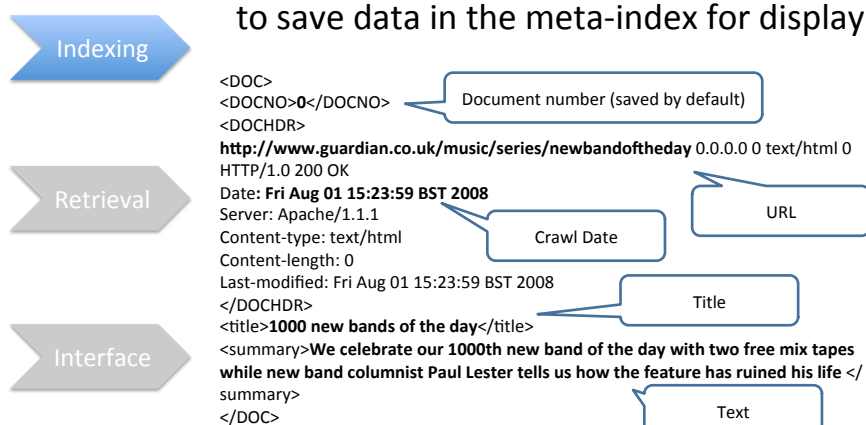
24/10/2011

© Terrier Team | University of Glasgow

87/134

Example Document

- The indexing process must be configured to save data in the meta-index for display



24/10/2011

© Terrier Team | University of Glasgow

88/134

Extract Headers

Header content can be saved using the TRECWebCollection class

Indexing

Retrieval

Interface

```
<DOC>
<DOCNO>0</DOCNO>
<DOCHDR>
http://www.guardian.co.uk/music/series/newbandoftheday 0.0.0.0 text/html 0
HTTP/1.0 200 OK
Date: Fri Aug 01 15:23:59 BST 2008
Server: Apache/1.1.1
Content-type: text/html
Content-
Last-mod
</DOCHDR>
<title>1000 new bands of the day</title>
<summary>We celebrate our 1000th new band of the day with two free mix tapes
while new band columnist Paul Lester tells us how the feature has ruined his life </
summary>
</DOC>
```

24/10/2011

© Terrier Team | University of Glasgow

89/134

Generate Abstracts

- Abstracts** from **tag contents** can be saved by TaggedDocument

Indexing

Retrieval

Interface

```
<DOC>
<DOCNO>0</DOCNO>
<DOCHDR>
http://www.guardian.co.uk/music/series/newbandoftheday 0.0.0.0 text/html 0
HTTP/1.0 200 OK
Date: Fri Aug 01 15:23:59 BST 2008
Server: Apache/1.1.1
Content-type: text/html
Content-
Last-modified: Fri Aug 01 15:23:59 BST 2008
</DOCHDR>
<title>1000 new bands of the day</title>
<summary>We celebrate our 1000th new band of the day with two free mix tapes
while new band columnist Paul Lester tells us how the feature has ruined his life </
summary>
</DOC>
```

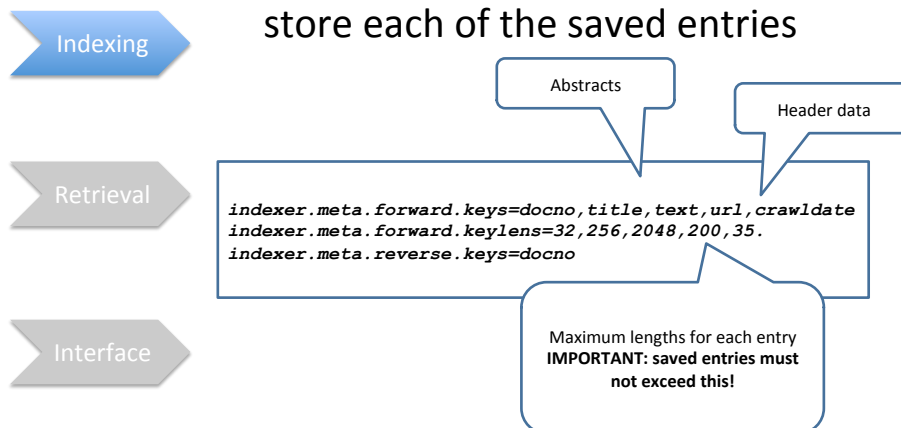
24/10/2011

© Terrier Team | University of Glasgow

90/134

Save to Meta Index

- The meta index needs to be told to store each of the saved entries



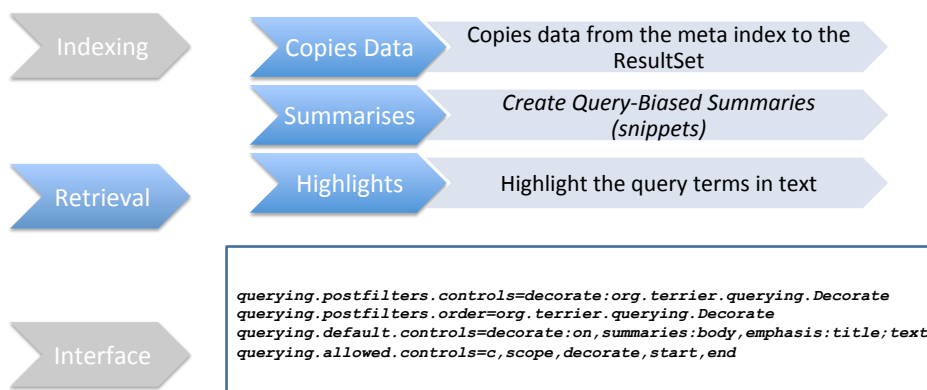
24/10/2011

© Terrier Team | University of Glasgow

91/134

Decorate ResultSet

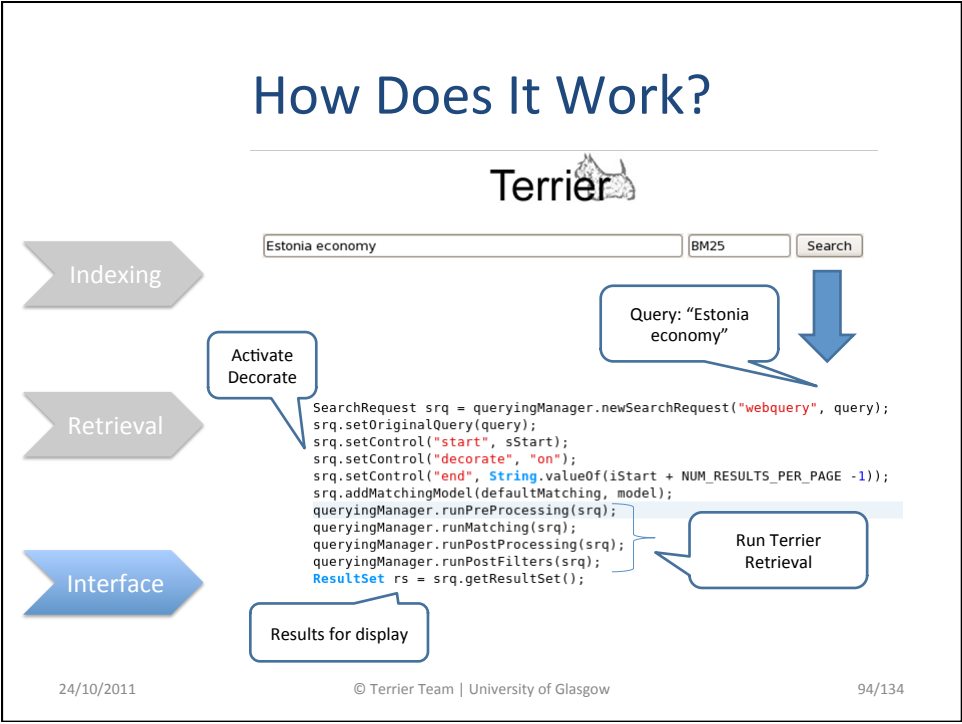
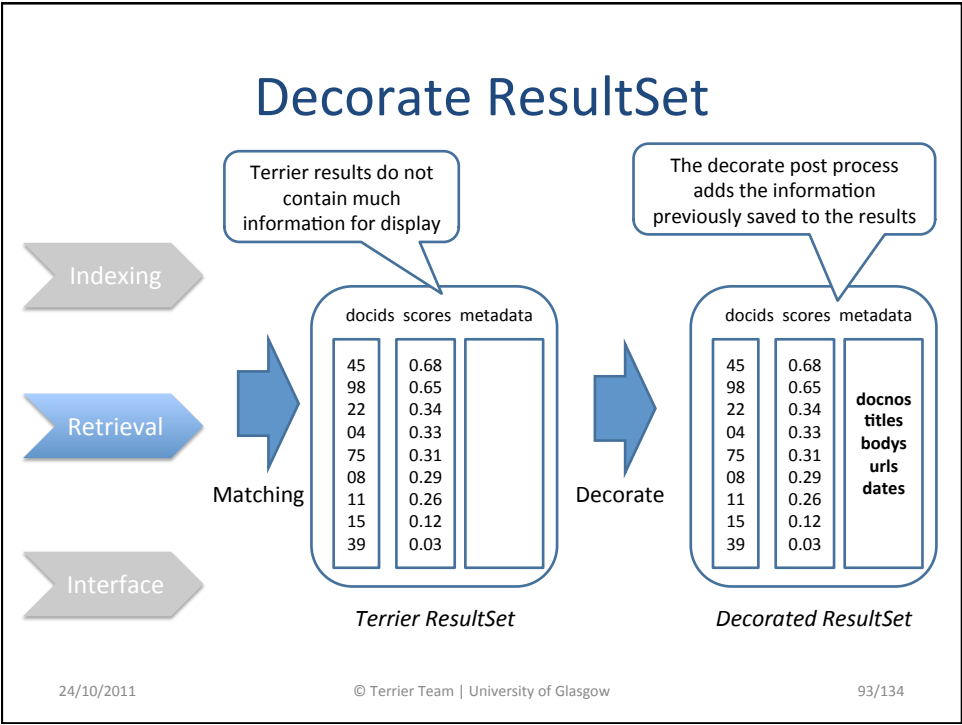
- Decorate



24/10/2011

© Terrier Team | University of Glasgow

92/134



Extending the Interface

- What if I want to display tweets?

– **twitter**

Search

– TREC Microblog Track

Terrier 

crowdsourcing				ConstantModel	C default	Search
Results for crowdsourcing, displaying 1-11 of 231 (Thu Oct 06 15:37:11 BST 2011)						
Rank	User	Tweet	Score			
1	 iMOHA	Crowdsourced Film A Hit At Sundance http://goo.gl/fb3Xpx	1.0			
2	 late2game	RT @rebeccacoleman: In case you haven't seen it yet, we now have our very own Canadian Crowdsourcing app: Fundchange! http://bit.ly/hcJQdN	1.0			
3	 amalucky	RT @Zee: Crowdsourcing : When, How and Who? http://bit.ly/gpK5Rd	1.0			

24/10/2011

© Terrier Team | University of Glasgow

95/134

Example JSON Tweet

```
{
  "id":28965131991912448,
  "id_str":"28965131991912448",
  "created_at":"Sun Jan 23 00:00:00 +0000 2011",
  "text":"@Biebercrombie haha I'm on my iPod and I'm on the tumblr app and it's working fine now;",
  "truncated":false,
  "retweet_count":0,
  "in_reply_to_screen_name":"Biebercrombie",
  "in_reply_to_status_id_str":"28964334474366976",
  "in_reply_to_status_id":28964334474366976,
  "contributors":null,
  "user":{
    "screen_name":"drizzybiebs",
    "protected":false,
    "lang":"en",
    "name":"neversayneverrrrÃ©Ã™Ã™",
    "profile_image_url":"http://a1.twimg.com/profile_images/
  },
  "entities":{
    "hashtags":[],
    "urls":[],
    "user_mentions":["Biebercrombie"]
  }
}
```

trec.collection.class=TwitterJSONCollection

Available from :
<http://ir.dcs.gla.ac.uk/wiki/Terrier/Tweets11>

24/10/2011

© Terrier Team | University of Glasgow

96/134

Configuration

```

trec.collection.class=TwitterJSONCollection

indexer.meta.forward.keys=docno,id,created_at,text,retweet_count,in_reply_to_screen_name,
in_reply_to_user_id,in_reply_to_status_id,user.name,user.screen_name,user.lang,user.profi
le_image_url,place.name,place.id,geo.lat,geo.lng,retweet.text,retweet.id,retweet.created_
at,retweet.retweet_count,retweet.in_reply_to_screen_name,retweet.in_reply_to_user_id,retw
eet.in_reply_to_status_id,retweet.user.name,retweet.user.screen_name,retweet.user.lang,re
tweet.user.profile_image_url,retweet.place.name,retweet.place.id,retweet.geo.lat,retweet.
geo.lng

indexer.meta.forward.keylens=32,30,30,200,10,30,30,30,60,30,10,250,160,30,30,30,200,30,30
,10,30,30,30,160,60,10,250,160,30,30,30

querying.postfilters.controls=decorate:org.terrier.querying.Decorate
querying.postfilters.order=org.terrier.querying.Decorate
querying.default.controls=decorate:on,emphasis:text
querying.allowed.controls=c,scope,decorate,start,end

```

Indexing collection format

These are all possible tweet fields

Max lengths for each field

Decorate the resultset during retrieval and highlight the query terms in the text (tweet)

24/10/2011
© Terrier Team | University of Glasgow
97/134

DEMO #4: Searching Tweets11

- Tweets11
 - Freely available
 - ~16million tweets

[\[Tweets11 Corpus\]](http://trec.nist.gov/data/tweets)
<http://trec.nist.gov/data/tweets>
- Used for the TREC Microblog Track

[\[TREC Microblog Track\]](https://sites.google.com/site/microblogtrack/)
<https://sites.google.com/site/microblogtrack/>

Large-Scale Information Retrieval Experimentation with Terrier

PART IV: EVALUATION (BATCH)

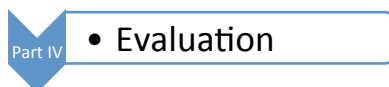
BY VASSILIS PLACHOURAS

24/10/2011

© Terrier Team | University of Glasgow

99/134

Tutorial Outline



Batch Evaluation

- Cranfield Evaluation
 - Current Instantiations
- Batch Retrieval
- Batch Evaluation
- Scripting Evaluation
- DEMO #5 (*work-along*)
 - Evaluating in batch mode

24/10/2011

© Terrier Team | University of Glasgow

100/134

IR System Evaluation

- What makes an IR system good?
 - Speed, size, quality of results
- Assessing the quality of results
 - Asking users directly
 - Measuring user engagement:
 - A/B testing
 - Split incoming user queries in two partitions
 - For each partition apply a different search algorithm
 - Measure changes in number of clicks, etc.
 - Standard test collections

24/10/2011

© Terrier Team | University of Glasgow

101/134

Cranfield Evaluation Paradigm

- Relative effectiveness comparison in controlled settings
- Standard test collection comprises:
 - Document set
 - Information needs
 - Relevance assessments

24/10/2011

© Terrier Team | University of Glasgow

102/134

Current Instantiations

- Text REtrieval Conference (TREC)
 - large-scale test collections
- Initiative for the Evaluation of XML Retrieval (INEX)
 - focus on structured and XML data
- NII Test Collection for IR Systems (NTCIR)
 - focus on cross language retrieval with English and Asian languages
- Cross Language Evaluation Forum (CLEF)
 - focus on cross language retrieval with European languages
- Forum for Information Retrieval Evaluation (FIRE)
 - focus on retrieval in English and South-Asian languages

24/10/2011

© Terrier Team | University of Glasgow

103/134

Batch Evaluation with Terrier

- Batch retrieval
`bin/trec_terrier.sh -r`
- Batch evaluation
 - To evaluate a single file
`bin/trec_terrier.sh -e filename.res`
 - To evaluate all files
`bin/trec_terrier.sh -e`

24/10/2011

© Terrier Team | University of Glasgow

104/134

DEMO #5: Retrieving in Batch Mode

- Setup index path

```
terrier.index.path=var/index
```

- Setup retrieval properties, for example

```
TrecQueryTags.doctag=TOP
```

```
TrecQueryTags.idtag=NUM
```

```
TrecQueryTags.process=TOP,NUM,TITLE
```

- Setup topics

```
trec.topics=etc/trcikm.topics
```

24/10/2011

© Terrier Team | University of Glasgow

105/134

DEMO #5: Retrieving in Batch Mode

- Run batch retrieval

```
bin/trec_terrier.sh -r -Dtrec.model=PL2
```

- Check output in folder `var/results`

- A result file

```
filename.res
```

- A file with corresponding retrieval settings

```
filename.res.settings
```

24/10/2011

© Terrier Team | University of Glasgow

106/134

DEMO #5: Evaluating in Batch Mode

- Setup qrels file
`trec.qrels=etc/trcikm.qrels`
- Run evaluation
`bin/trec_terrier.sh -e`
- Results are stored in folder
`var/results/*.eval`

24/10/2011

© Terrier Team | University of Glasgow

107/134

DEMO #5: Evaluating in Batch Mode

- Large scale experimentation
 - Using scripts to run a series of experiments
- Run experiments with different models

```
#!/bin/bash
#batch retrieval
for m in PL2 DPH BM25; do
  bin/trec_terrier.sh -r -Dtrec.model=$m -Dtrec.runtag=$m \
    -Dtrec.results.file=var/results/$m.res;
done

#batch evaluation
bin/trec_terrier.sh -e -Dtrec.qrels=etc/trcikm.qrels
```

24/10/2011

© Terrier Team | University of Glasgow

108/134

Large-Scale Information Retrieval Experimentation with Terrier

PART IV: EVALUATION (CROWDSOURCING)

BY RICHARD MCCREADIE

24/10/2011

© Terrier Team | University of Glasgow

109/134

Tutorial Outline



• Evaluation

Extending Evaluation

- Crowdsourcing
- MechanicalTurk
- Terrier and Crowdsourcing
 - Converting a ResultSet to a HIT
 - From a Web Interface to an HTML form
 - Validation and Consensus
- DEMO #6
 - Crowdsourcing Relevance Judgements for the TREC 2011 Crowdsourcing Track

24/10/2011

© Terrier Team | University of Glasgow

110/134

Extending Terrier

- So far we have covered Terrier functionality provided in the open source release
- In this last section of the tutorial, we give an example of how to add new functionality to Terrier

Integrate crowdsourcing for relevance assessment

[Alonso et al., SIGIR Forum 2008]

24/10/2011

© Terrier Team | University of Glasgow

111/134

Motivation

- What if I have a new task/document collection that I want to evaluate upon?
 - I need to assess the documents returned by Terrier for relevance
- I could:
 - Propose a new TREC track
 - Long term, difficult
 - Assess documents manually
 - Fast for small collections, non-scalable, boring
 - Pay students to do it
 - Costly, recruitment can be difficult
 - **Crowdsource document assessment**
 - **Fast, cheap, scalable**



24/10/2011

© Terrier Team | University of Glasgow

112/134

What is Crowdsourcing

- Outsourcing of tasks normally done in-house to a wider group (the crowd) with an open call

[Howe, 2009]

- Positive
 - Fast, cheap, scalable, on-demand
- Negative
 - Setup, quality

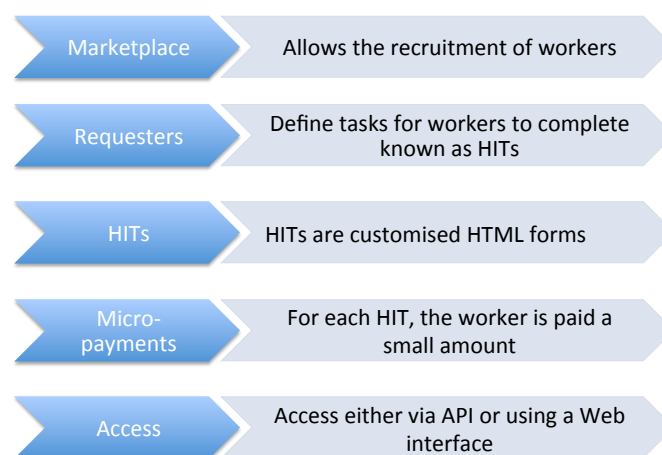


24/10/2011

© Terrier Team | University of Glasgow

113/134

amazon mechanical turk™ Artificial Artificial Intelligence



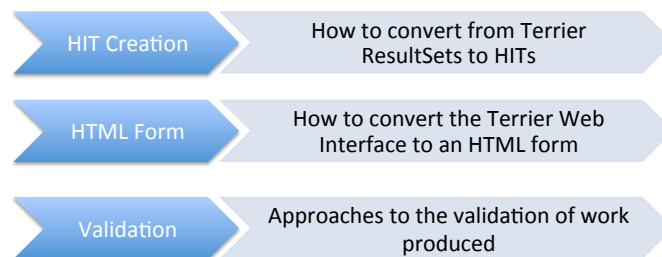
24/10/2011

© Terrier Team | University of Glasgow

114/134

Aim

- We will show how to integrate Mturk with Terrier for automatic generation of relevance assessments

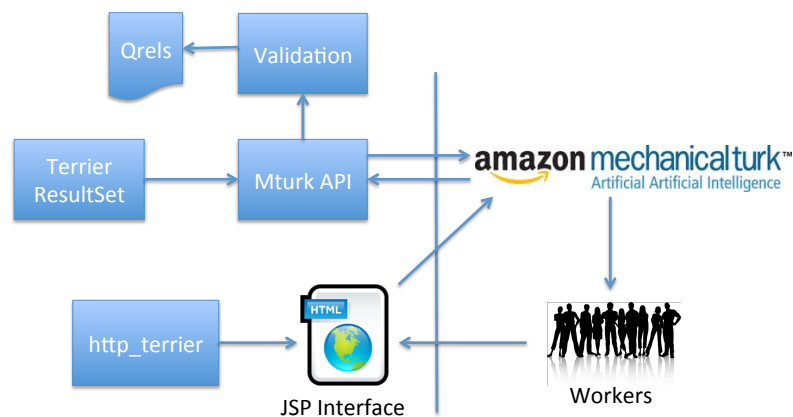


24/10/2011

© Terrier Team | University of Glasgow

115/134

Overview



24/10/2011

© Terrier Team | University of Glasgow

116/134

Converting a Terrier ResultSet into HIT

Parameters that are passed to the Web interface

XML Schema location

Where the Terrier Web interface hosted?

How big should Mturk make the Iframe that the Web interface loads in?

docids scores metadata

45	0.68	
98	0.65	
22	0.34	
04	0.33	
75	0.31	
08	0.29	
11	0.26	
15	0.12	
39	0.03	

Decorated ResultSet

XML Header

<ExternalQuestion xmlns=\http://mechanicalturk.am ... \>

<ExternalURL>\http://demos.terrier.org/crowdsourcing/results.jsp?query=Query& ... </ExternalURL>

<FrameHeight>1000</FrameHeight>

</ExternalQuestion>

24/10/2011

© Terrier Team | University of Glasgow

117/134

Web Interface to HTML Form

Worker instructions

Fixed query and description

HTML form questions

Terrier

Estonia economy BM25 Search

Results for Estonia economy, displaying 1-11 of 18047

Rank	Document	Score
1	http://server.soros.org/80/estonia/estocoun.html	11.148
	Open Estonia Foundation - About the Country	
	server.soros.org/80/estonia/estocoun.html - WT09-B18-220 - 853062034000	
	Significance of the Estonian-Russian Interest Rate Differential	
	There are few better proxies for measuring economic and political risk in the former Soviet Union than the interest rate differential between Estonia and Russia	11.046
	www.kemper.com/80/ite/cutious/global_economy/weekly_report/estonia1.html - WT01-B13-137 - 852570874000	

+ Instructions (Click to Expand)

Hopi food Search

Description: What kinds of food do Hopis eat?

Results for Hopi food, displaying 1-5 of 5

Rank	Document	Relevance?
1	Hopi Indians	☐ Highly Relevant ☐ Relevant ☐ Not Relevant
	Corn relies on the farmer to survive, and the Hopi relies on the corn - all life is designed to be interrelated. The Hopi Indians grow food similar to the Navajo Indians	
	www.cystalks.com/hopi1.html - claweb09-en0006-17-27544	
2	Learn about the History of the Hopi Indians	☐ Highly Relevant ☐ Relevant ☐ Not Relevant
	Squash, beans, pumpkins, and other various fruits were also grown and eaten. The women of the Hopi tribe were designated to take care of the home, look after the children, and cook	
	www.indians.org/articles/hopi-indians.html - claweb09-en00510-44-22927	

24/10/2011

© Terrier Team | University of Glasgow

118/134

59

Web Interface to HTML Form

Recall:
`<ExternalURL>http://demos.terrier.org/
crowdsourcing/results.jsp?query=Query& ... </
ExternalURL>`

```
SearchRequest srq = queryingManager.newSearchRequest("webquery", query);
srq.setOriginalQuery(query);
srq.setControl("start", sStart);
srq.setControl("decorate", "on");
srq.setControl("end", String.valueOf(iStart + NUM_RESULTS_PER_PAGE - 1));
srq.addMatchingModel(defaultMatching, model);
queryingManager.runPreProcessing(srq);
queryingManager.runMatching(srq);
queryingManager.runPostProcessing(srq);
queryingManager.runPostFilters(srq);
ResultSet rs = srq.getResultSet();
```

Same results as when
producing the HIT

```
Map<String,String> queryDescriptions = queryLoader.loadDescriptions("/queries/descriptions.txt");
```

Load the query
descriptions from a file

24/10/2011

© Terrier Team | University of Glasgow

119/134

Web Interface to HTML Form

Form submit
address to Mturk

This function check to see if all
judgments have been done yet
for this worker

```
out.print("<form id=\"relevanceform\" action=\"http://www.mturk.com/mturk/externalSubmit\">");
out.print("<table cellpadding=\"4\" cellspacing=\"0\" border=\"0\">");
out.print("<tbody>");
out.print("<tr>");
out.print("<td valign=\"center\"><input type=\"radio\" onclick='checkAllJudgmentsMade(\"+i+\");' value=\"1\">");
out.print("<td><span class=\"answertext\" id=\"A\"+i+\"2\">Highly Relevant</span></td>");
out.print("</tr>");
out.print("<tr>");
out.print("<td valign=\"center\"><input type=\"radio\" onclick='checkAllJudgmentsMade(\"+i+\");' value=\"1\">");
out.print("<td><span class=\"answertext\" id=\"A\"+i+\"1\">Relevant</span></td>");
out.print("</tr>");
out.print("<tr>");
out.print("<td valign=\"center\"><input type=\"radio\" onclick='checkAllJudgmentsMade(\"+i+\");' value=\"0\">");
out.print("<td><span class=\"answertext\" id=\"A\"+i+\"0\">Not Relevant</span></td>");
out.print("</tr>");
out.print("</tbody>");
out.print("</table>");
out.print("<input type=\"hidden\" name=\"hitid\" id=\"hitid\" value=\"\"+MTurkHITID+\"\" />");
out.print("<input type=\"hidden\" name=\"assignmentId\" id=\"MTurkAssignmentID\" value=\"\"+MTurkHITID+\"\" />");
```

Pass back parameters to
Mturk so it knows which HIT
and assignment this is

Relevance
assessment radio
buttons

24/10/2011

© Terrier Team | University of Glasgow

120/134

Work Validation

- Work needs to be validated
 - Malicious workers
 - Bots
- Evaluate work based upon
 - Inter-worker agreement
 - Heuristics, e.g. Time taken, patterns
 - Gold judgments



Is my work valid?

Crowdsourced TREC Relevance Evaluation

Workers:

WorkerID	#	Acc	Impact	AvgTime	Notes
A1ITQ1GSWH5WTH	10	50.0%	0.327%	17	
ARUM0000AA2CF	10	30.0%	0.327%	16	
A3Q77NZ67J30S	65	60.0%	0.163%	11	
A2EZY98YRIWC86	365	40.0%	11.94%	15	
A3B5LEX8EGBPP8	10	70.0%	0.327%	15	

24/10/2011

© Terrier Team | University of Glasgow

121/134

Consensus

- Often you will have many worker judge each document to increase quality
 - Known as redundant judging
- Use majority voting to merge multiple assessments to a single label



[Snow et al., EMNLP 2008]
[Callison-Burch et al., EMNLP 2009]

24/10/2011

© Terrier Team | University of Glasgow

122/134

DEMO #6: Crowdsourcing Relevance Assessments

- Crowdsource relevance assessments
 - TREC 2011 Crowdsourcing Track
- [TREC Crowdsourcing track
https://sites.google.com/site/treccrowd2011/]

+ Instructions (Click to Expand)

Description: What kinds of food do Hopis eat?

Results for Hopi food, displayed 1-5 of 5

Rank	Document	Relevance?
1	Hopi Indians Corn relies on the former to survive, and the Hopi relies on the corn - all life is designed to be interrelated. The Hopi Indians grow food similar to the Navajo Indians. www.cystadinks.com/hopi1.htm - clusweb009-en0006-17-27544	☐ Highly Relevant ☐ Relevant ☐ Not Relevant
2	Learn about the history of the Hopi Indians Squash, beans, pumpkins, and other various fruits were also grown and eaten. The women of the Hopi tribe were designated to take care of the home, look after the children, and cook. www.indians.org/articles/hopi-indians.html - clusweb009-en0010-44-22927	☐ Highly Relevant ☐ Relevant ☐ Not Relevant

Large-Scale Information Retrieval Experimentation with Terrier

WRAPPING UP

BY RODRYGO SANTOS

Achieved Outcomes

- Large-scale indexing
 - Index to support different requirements
 - Scale indexing using MapReduce
- Large-scale retrieval
 - Leverage the richest open-source retrieval API
 - Extend retrieval for non-standard tasks
- Large-scale evaluation
 - Evaluate multiple approaches in batch mode
 - Extend evaluation through crowdsourcing

24/10/2011

© Terrier Team | University of Glasgow

125/134

Summary

- Terrier is ideal for...
 - ... handling large corpora
 - ... tackling complex search tasks
 - ... evaluating multiple experimental scenarios
- Open-source IR platform since 2004
 - Currently in version 3.5

Contributions are always welcomed!

24/10/2011

© Terrier Team | University of Glasgow

126/134

Useful Links

- Terrier Website
– <http://terrier.org>
- Terrier Documentation
– <http://terrier.org/docs/v3.5>
- Terrier Forum
– <http://terrier.org/forum>
- Terrier Issue Tracker
– <http://terrier.org/issues/browse/TR>

24/10/2011

© Terrier Team | University of Glasgow

127/134

Cite Terrier in Your Research

- [Ounis et al., OSIR 2006] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In Proc. of the OSIR Workshop at SIGIR, 2006
- [Ounis et al., Novatica 2007] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in Terrier: a search engine for advanced retrieval on the Web. In Novatica/UPGRADE Special Issue on Next Generation Web Search, 2007
- [Ounis et al., ECIR 2005] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson. Terrier information retrieval platform. In Proc. of ECIR, 2005

Check out all Terrier-related publications at:
<http://terrier.org/publications.html>

24/10/2011

© Terrier Team | University of Glasgow

128/134

Other References

- [Alonso et al., SIGIR Forum 2008] O. Alonso. Crowdsourcing for relevance evaluation. SIGIR Forum, 2008
- [Amati, PhD Thesis 2003] Probability Models for Information Retrieval based on Divergence From Randomness. PhD Thesis, 2003
- [Amati et al., ECIR 2006] G. Amati. Frequentist and Bayesian approach to information retrieval. In Proc. of ECIR, 2006
- [Amati et al., TREC 2007] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog track. In Proc. of TREC, 2007
- [Callison-Burch, EMNLP 2009] C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In Proc. of EMNLP, 2009
- [Clinchant & Gaussier, SIGIR 2010] S. Clinchant and É. Gaussier. Information-based models for adhoc IR. In Proc. of SIGIR, 2010

24/10/2011

© Terrier Team | University of Glasgow

129/134

Other References

- [Dean & Ghemawat, OSDI 2004] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In Proc. of OSDI, 2004
- [Dincer et al., TREC 2009] B. T. Dincer, I. Kocabaş, and B. Karoğlu. IRRA at TREC 2009: Index term weighting based on Divergence From Independence model. In Proc. of TREC, 2009
- [Gil-Costa et al., SPIRE 2011] V. Gil-Costa, R. L. T. Santos, C. Macdonald, and I. Ounis. Sparse spatial selection for novelty-based search result diversification. In Proc. of SPIRE, 2011
- [Howe, 2009] J. Howe. The rise of Crowdsourcing, 2010
<http://www.wired.com/wired/archive/14.06/crowds.html>
- [Macdonald et al., CLEF 2005] C. Macdonald, V. Plachouras, B. He, C. Lioma, and I. Ounis. University of Glasgow at WebCLEF 2005: Experiments in per-field normalisation and language specific stemming. In Proc. of CLEF, 2005

24/10/2011

© Terrier Team | University of Glasgow

130/134

Other References

- [\[Macdonald et al., TOIS 2011\]](#) C. Macdonald, I. Ounis, and N. Tonellotto. Upper bound approximations for dynamic pruning. ACM Trans. Inf. Syst., 2011
- [\[McCreadie et al., CSDM 2011\]](#) R. McCreadie, C. Macdonald, and I. Ounis: Crowdsourcing Blog track top news judgments at TREC. In Proc. of the CSDM Workshop at WSDM, 2011
- [\[McCreadie et al., IPM 2011\]](#) R. McCreadie, C. Macdonald, and I. Ounis. MapReduce indexing strategies: Studying scalability and efficiency. Inf. Proc. and Manage., 2011
- [\[McCreadie et al., LSDSIR 2009\]](#) R. McCreadie, C. Macdonald, and I. Ounis. Comparing distributed indexing: To MapReduce or not? In Proc. of the LSDSIR Workshop at SIGIR, 2009
- [\[McCreadie et al., SIGIR 2009\]](#) R. McCreadie, C. Macdonald, and I. Ounis. On single-pass indexing with MapReduce. In Proc. of SIGIR, 2009

24/10/2011

© Terrier Team | University of Glasgow

131/134

Other References

- [\[Metzler & Croft, SIGIR 2005\]](#) D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In Proc. of SIGIR, 2005
- [\[Peng et al., SIGIR 2007\]](#) J. Peng, C. Macdonald, B. He, V. Plachouras, and I. Ounis. Incorporating term dependency in the DFR framework. In Proc. of SIGIR, 2007
- [\[Plachouras & Ounis, ECIR 2007\]](#) V. Plachouras and I. Ounis. Multinomial randomness models for retrieval with document fields. In Proc. of ECIR, 2007
- [\[Santos et al., ECIR 2009\]](#) R. L. T. Santos, B. He, C. Macdonald, and I. Ounis. Integrating proximity to subjective sentences for blog opinion retrieval. In Proc. of ECIR, 2009
- [\[Snow et al., EMNLP 2008\]](#) R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In Proc. of EMNLP, 2008

24/10/2011

© Terrier Team | University of Glasgow

132/134

Other References

- [\[Tombros & Sanderson, SIGIR 1998\]](#) A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In Proc. of SIGIR, 1998
- [\[Zaragoza et al., TREC 2004\]](#) H. Zaragoza, N. Craswell, M. J. Taylor, S. Saria, and S. E. Robertson. Microsoft Cambridge at TREC 13: Web and Hard tracks. In Proc. of TREC, 2004

Questions?

