Terrier Information Retrieval Platform

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He,
Craig Macdonald and Douglas Johnson

Department of Computing Science
University of Glasgow
ounis, gianni, vassilis, ben, craigm, johnsoda @dcs.gla.ac.uk

Terrier

http://ir.dcs.gla.ac.uk/terrier/

The Information Retrieval Group http://ir.dcs.gla.ac.uk

Overview

Terrier is an open-source Information Retrieval (IR) platform for the development of retrieval applications and the experimentation with large-scale document collections

1. About Terrier

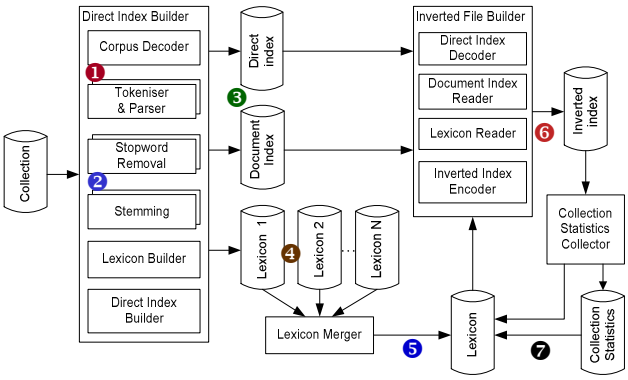
- **Terrier** (Terabyte Retriever) is a **test-bed** for the rapid development of large-scale retrieval applications
- Terrier is based on the highly effective **parameter-free** probabilistic weighting models of the **Divergence From Randomness** (**DFR**) framework
- It includes various other well-known IR models, such as tf-idf, BM25 and language modelling, for cross-comparison
- Terrier allows to assess the current state-of-the-art IR models and to quickly experiment with new concepts/ideas in various settings
- A core version of Terrier is available as open source software

2. Data Structures

Lexicon (10.5% of index)	**Term**, **Term id**, **Document frequency**, **Term Frequency**, **End byte offset in inverted file**, **End bit offset in inverted file**
Lexicon Index	**Offset of an entry in the lexicon**
Document Index (5% of index)	**Document id**, **Document Length**, **Document number**, **End byte offset in direct file**, **End bit offset in direct file**
Direct Index (40% of index)	**Term id gap** (gamma code), **Term frequency** (unary code), **Fields**, **Block frequency** (unary code), **[Block id gap** (gamma code)**]**
Inverted Index (44.5% of index)	**Document id gap** (gamma code), **Term frequency** (unary code), **Fields**, **Block frequency** (unary code), **[Block id gap** (gamma code)**]**
Collection Statistics	**# of documents, # of tokens, # of unique terms, # of pointers**

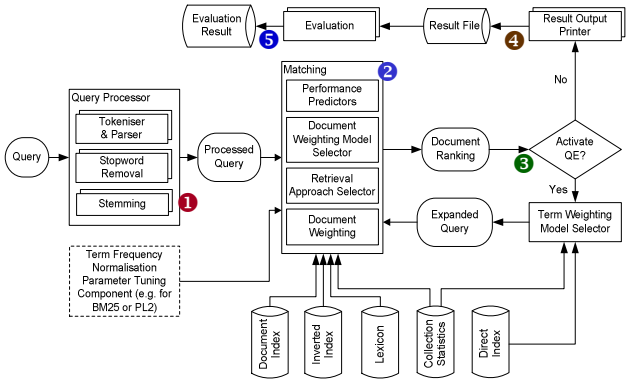
3. Indexing

1. Tokenise & parse documents
2. Application-dependent stopword removal and stemming
3. Build the direct and document indices
4. Use less memory by building lexicons for parts of the collection
5. Merge temporary lexicons to form one lexicon
6. Build the inverted index from the existing data structures
7. Update collection statistics and lexicon



4. Querying

1. Application-dependent query processing
2. Select optimal weighting model and/or appropriate retrieval approaches per query
3. If query expansion is activated, select a term weighting model
4. Application-dependent result rendering, e.g. XML
5. Standard evaluation techniques



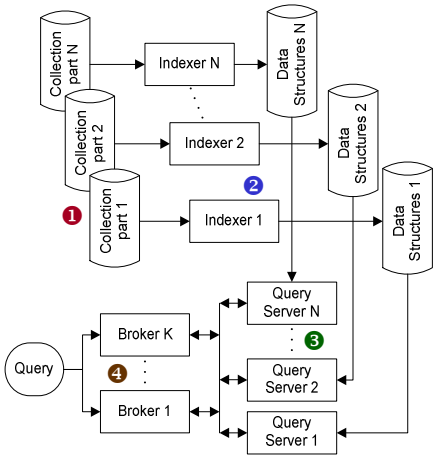
5. Large-scale collections

1. Split the collection into disjoint parts
2. Index the parts independently
3. Use a query server for each part
4. Brokers dispatch queries and merges results

Indexing and Retrieval from .GOV2 (426GB)

Data Structures	Size	% of .GOV2
All structures	17.48GB	4.1%
Inverted files	7.77GB	1.8%
Direct files	7.00GB	1.6%
Lexicons	1.84GB	0.4%
Document indices	0.87GB	0.2%

Run Description	MAP	bpref	P10
Median	0.1427	0.2015	0.4102
PL2+short queries	0.2709	0.3026	0.5306
PL2+short queries+anchors	0.2690	0.3025	0.5245
PL2+long queries	0.3054	0.3356	0.6163
PL2+long queries+QE	0.3075	0.3359	0.6327



6. Terrier Retrieval Features

Terrier has various useful features for enhancing retrieval effectiveness:

- Pre-retrieval query performance predictors
  →Marginal computational overhead
- Selection of optimal weighting model
  →Based on the statistical features of the query
- Selection of retrieval approaches on a per-query basis
  →Based on features of the set of retrieved documents
- Parameter-free automatic query expansion
  →The weights of expanded terms depends on the terms' statistics
- Possibility to tune the term frequency normalisation, e.g. for BM25 or PL2
  →Very robust and effective

7. What we have learnt?

- Developed a cutting-edge technology from a laboratory setting to a product release
- Building a sustainable development platform is equally important as creating new IR models
- Our objective to generate parameter-free models has given us more insight about IR
- New ideas can be easily implemented and evaluated

This work is funded by a UK EPSRC project grant, number GR/R90543/01. The project funds the development of Terrier IR framework (http://ir.dcs.gla.ac.uk/terrier)

EPSRC Engineering and Physical Sciences Research Council