Terrier: TERABYTE RETRIEVER

Iadh Ounis

Information Retrieval Group

About Terrier (1)

- Terrier is a modular platform for the rapid development of large-scale Information Retrieval (IR) applications.
- Terrier is based on a new parameter-free probabilistic framework for IR (DFR), allowing adaptable term weighting functionalities
- Terrier includes state-of-the-art functionalities such as:
  - hyperlink structure analysis,
  - combination of evidence approaches,
  - automatic query expansion/re-formulation techniques,
  - query performance predictors
  - compression techniques.
- It is written in Java (and Perl)

About Terrier (2)

- Since 2001, Terrier was supported by a 30- months EPSRC grant
- Currently 3 researchers, 5 PhD students and 5 programmers constitute the Terrier team
- Terrier deploys over 50 term weighting/ matching functions, including various DFR models, and the well-established BM25 and language modelling approaches
- Terrier has a comprehensive documentation
- Terrier has a robust and effective crawler, called Labrador

About Terrier (3)

- Terrier allows a large-scale experimentation conducted in a robust, transparent, reproducible, modular, platform-independent, and without constraints and parameters
- Terrier allowed us to easily assess and improve IR technology
- Terrier allowed the rapid experimentation of new concepts/ideas on various collections, and in different settings
- Terrier has been successfully used for various retrieval tasks, in a centralised or distributed setting.

The Divergence from Randomness Framework (DFR)

- The DFR approach is based on a simple idea:
  - "*The more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the word t in the document d*"

- The DFR approach can be defined as the divergence of two probabilities measuring the amount of randomness of two different sources of evidence. (See Gianni Amati's thesis, 2003)

Terrier

UNIVERSITY of GLASGOW

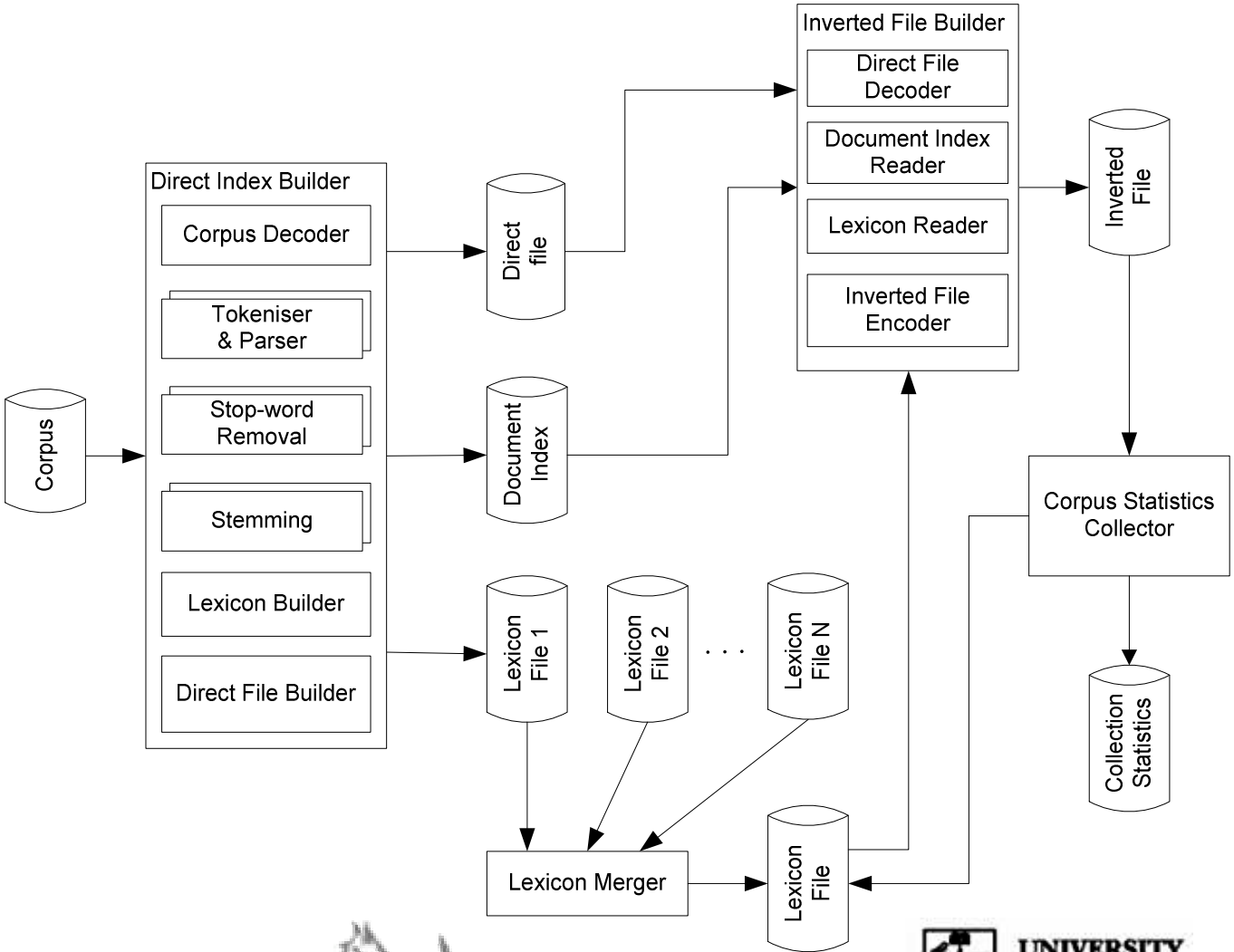
The DFR framework

- The use of parameter-free probabilistic models is new in the IR field
- Unlike Okapi's BM25 and the language modelling approaches, the DFR framework offers parameter-free baselines
  - Terrier provides methods for setting automatically the efficiency parameters
    - Length normalisation, query reformulation
- Terrier learns from empirical data and adapts to the users' information needs and queries
- Terrier has an outstanding performance with respect to other current public technologies on various large-scale collections and different retrieval tasks

Terrier

UNIVERSITY of GLASGOW

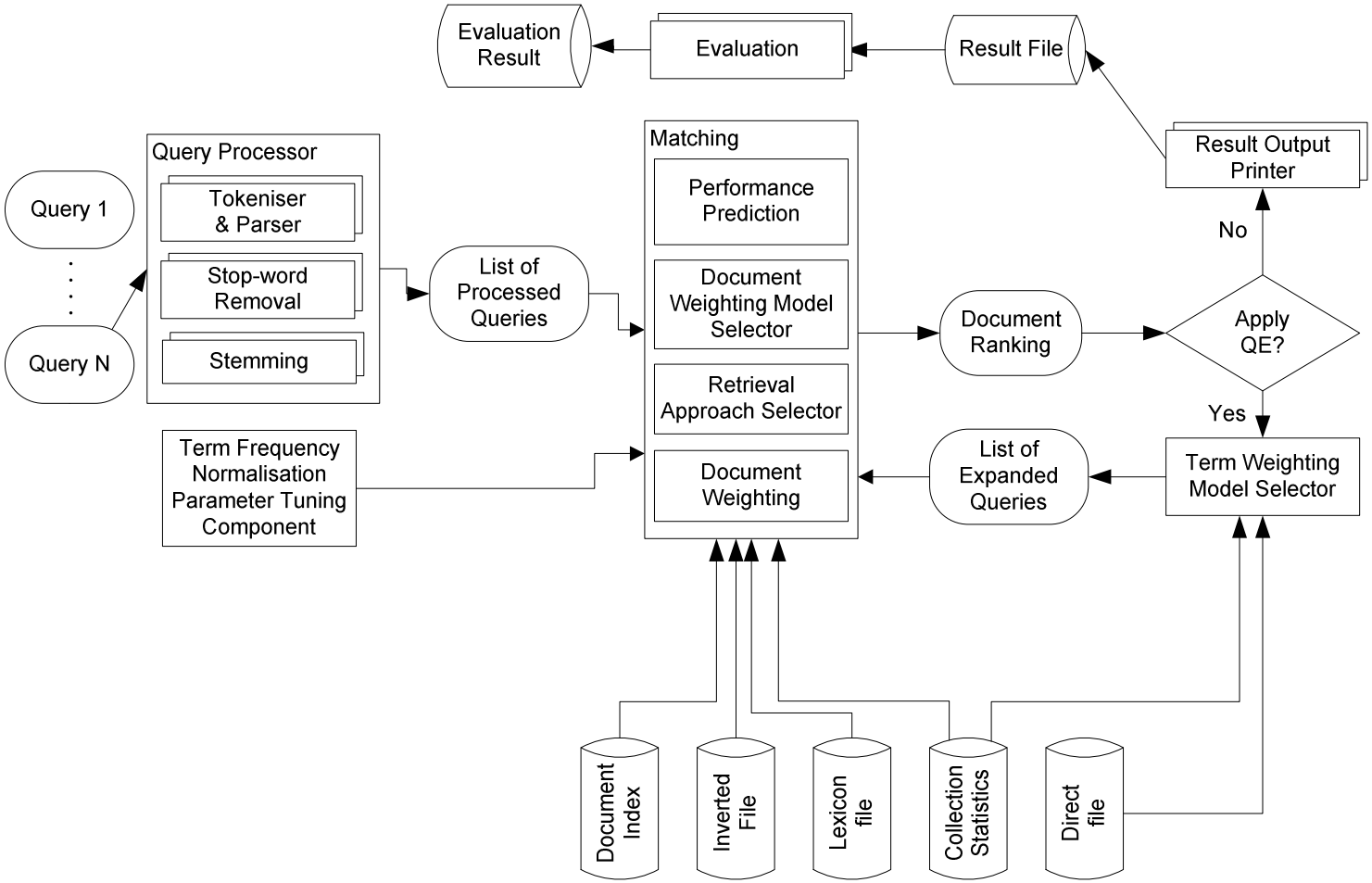
Terrier Indexing



Terrier

UNIVERSITY of GLASGOW

Terrier Querying



Terrier

UNIVERSITY of GLASGOW

Some Figures (1)

- The .GOV2 collection with light porter stemmer
  - Also called the Terabyte TREC collection
  - 25,205,179 documents
  - Average size of a document: ~17.7 kb

total size of index files on disk: 16.87 GB

inverted files: 7.82 GB

direct files: 7.06 GB

time to build : 3 days (2 processors)

time to retrieve: 4sec/query (8 processors)

A much better throughput can be achieved using more processors.

Terrier

UNIVERSITY of GLASGOW

Some Figures (2)

- The .GOV with anchor text, titles and headings
  - Also called the 18GB TREC collection
  - 1,247,753 documents

total size of index files on disk: 1.14 GB
inverted file: 510MB
direct file: 497MB
time to build : 14.1 hours
time to retrieve : 0.8sec/query
number of processors: 1

Some Figures (3)

- Crawl of the GLA domain
  - Number of documents: 353,752

total size of index files: 206 MB
inverted file: 71 MB
direct file: 70 MB
time to build : 57 minutes
Querying time : ~1 sec/query
number of processors: 1

- Crawl of the DCS domain
  - Number of documents: 49,115

total size of index files: 40 MB
inverted file: 8.3 MB
direct file: 9.1 MB
time to build : 13 minutes
Querying time : <1 sec/query
number of processors: 1

UNIVERSITY of GLASGOW

Web Hyperlink Analysis

- Most current commercial search engines incorporate a link analysis component in their document ranking mechanism.
  - e.g. Google's PageRank or Kleinberg's HITS algorithm

- Terrier includes a novel link analysis component
  - More general than Google's PageRank
  - No use of parameters such as the damping factor
  - It can be applied in a query-dependent or independent way
  - Could be used in various applications, e.g. multilingual retrieval

Terrier

UNIVERSITY of GLASGOW

Other Features

- Length normalisation (Ben He's thesis)
  - Collection-independent
    - Assume a constant optimal normalisation effect over collections
    - For a given collection/query, apply the parameter such that it gives an optimal perf0r
- Retrieval approaches selector (Ben He's thesis)
  - Apply the optimal matching functions/ QE models on a per-query basis
  - Involves a query clustering process allocating the optimal weighting models, including document ranking and query expansion models, on a per-query basis
- Dynamic selection of Web retrieval approaches (Vassilis Plachouras's thesis)
  - Employs evidence from the *textual content*, *URL type*, and the *hyperlink structure* of the set of retrieved documents to optimally select the appropriate retrieval approach(es) on a per-query basis.
  - Fully automatic decision mechanism
- **Terrier has a set of performance predictors**
  - State-of-the-art predictors
- A very low computational overhead

Terrier

UNIVERSITY of GLASGOW

Applications Based on Terrier

- TREC Conference Proceedings 2001-2004
  - Adhoc retrieval tasks, Web retrieval tasks, robust retrieval tasks.

- Multilingual retrieval
  - French retrieval (See CLEF 2004)
  - Italian retrieval (See CLEF 2003)
  - Cross-language retrieval
  - Use of DFR has gained an increasing attention in 2004

- Intranet Search
  - DCS & GLA search facility
  - Italian Ministry of Communication

Terrier

UNIVERSITY of GLASGOW