
Open Source IR Tools and Libraries

Giorgos Vasiliadis, gvasil@csd.uoc.gr

CS-463 Information Retrieval Models

Computer Science Department

University of Crete

Outline

- Google Search API



- Lucene



- Terrier



- Lemur



Google Search API



Google Search API: Overview

- The API exposes the Google engine to developers.
 - You can write scripts that access the Google search in real-time.
- Google no longer issuing new API keys for the SOAP Search API.
- Instead, Google provides an AJAX Search API.
 - You can put Google Search in your web pages with JavaScript.

Google Search API: SOAP

- Based on the Web Services Technology SOAP (the XML-based Simple Object Access Protocol).
- Developers write software programs that connect remotely to the Google SOAP Search API service.
- Developers can issue search requests to Google's index of billions of web pages and receive results as structured data, access information in the Google cache and check the spelling of words.
- Limitations
 - Default limit of 1,000 queries per day.
 - Can only query for 10 results a time
 - Can only access Google Web Search (not Google Images, Google Groups and so on).

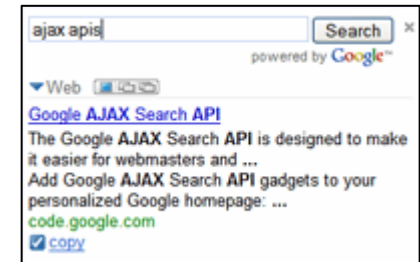
Google Search API: AJAX

- Lets you put Google Search in your web pages with JavaScript.
- Does not have a limit on the number of queries per day.
- Supports additional features like Video, News, Maps, and Blog search results.

Google Search API: AJAX

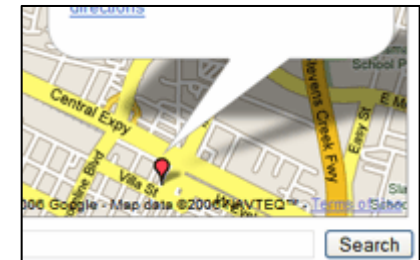
■ Web Search

- Incorporate results from [Web Search](#), [News Search](#), and [Blog Search](#)



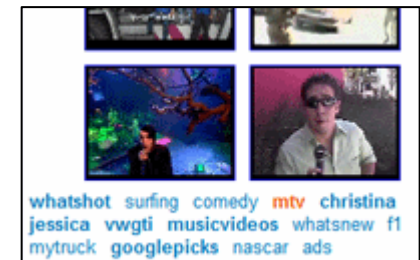
■ Local Search

- Provides access to local search results from [Google Maps](#).



■ Video Search

- Incorporate a simple search box
- incorporate dynamic, search powered strips of video and book thumbnails.



Google Search API: Demo

The screenshot shows a web browser window with two search results side-by-side. The left search is for 'VW GTI' and the right is for 'Ferrari Enzo'. Both are powered by Google. The 'VW GTI' results include links to 'GTI Mk V Home', 'Volkswagen of America', '2007 Volkswagen GTI 5-door - Short Take Road Tests - Car and ...', 'Volkswagen Golf - Wikipedia, the free encyclopedia', 'VW GTI Campaign', and 'VW GTI - 2006 Volkswagen GTI Road Test Review'. The 'Ferrari Enzo' results include video thumbnails and links to 'the FERRARI ENZO', 'Eddie Griffin ferrari Enzo Crash', 'REDLINE - Eddie Griffin Crashes Ferrari Enzo', 'Fifth Gear - Ferrari Enzo vs McLaren F1', 'ferrari enzo autopista costanera norte santiago', 'Eddie Griffin crashed a rare Ferrari Enzo worth \$1.5 million into a concrete barrier while', and 'Ferrari Enzo crash'.

Small Results, Web Large

VW GTI Search powered by Google™

Video Web Blog

[GTI Mk V Home](#)
The portal to everything GTI Mk V - features and specs, photo gallery, special offers - you can even build your own GTI Mk V and take it for a Joyride from ...
www.vw.com

[Volkswagen of America](#)
Volkswagen of America presents US vehicle information, pricing, incentives, deals, comparisons on Eos, GTI, Jetta, New Beetle, New Beetle Convertible, ...
www.vw.com

[2007 Volkswagen GTI 5-door - Short Take Road Tests - Car and ...](#)
Performance met utility, fell in love, and out came a turbocharged five-door.,
www.caranddriver.com

[Volkswagen Golf - Wikipedia, the free encyclopedia](#)
The Volkswagen Rabbit GTI, the North American version of the high-performance Golf GTI, debuted in Canada in 1979 and in the United States for 1983 model ...
en.wikipedia.org

[VW GTI Campaign](#)
20 years ago Volkswagen created a monster. In 1984, VW introduced the GTI, giving birth to Hot Hatch culture, and along with it came a new breed of car ...
www.cpbgroup.com

[VW GTI - 2006 Volkswagen GTI Road Test Review](#)
Read the latest Volkswagne GTI review by seasoned automotive journalists.
www.automedia.com

[Build a GTI Mk V](#)

All Large

Ferrari Enzo Search powered by Google™

Video Web Blog

[the FERRARI ENZO](#)
THE WORLD'S FASTEST CAR... NOW on YOUTUBE.... MY GOD! THAT'S A HARD
Apr 17, 2006
www.youtube.com

[Eddie Griffin ferrari Enzo Crash](#)
Eddie Griffin Enzo Crash Caught on Tape LIVE!
Mar 27, 2007
www.youtube.com

[REDLINE - Eddie Griffin Crashes Ferrari Enzo](#)
<http://www.247TVshows.com> Eddie Griffin Crashes Ferrari on set of Redline Movie
Mar 28, 2007
www.youtube.com

[Fifth Gear - Ferrari Enzo vs McLaren F1](#)
Fifth Gear - Ferrari Enzo vs McLaren F1
Jul 27, 2006
www.youtube.com

[ferrari enzo autopista costanera norte santiago](#)
A ferrari enzo racing whit a toyota supra and a subaru impreza wrx in costanera norte highway at
Feb 04, 2006
video.google.com

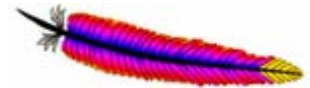
[Eddie Griffin crashed a rare Ferrari Enzo worth \\$1.5 million into a concrete barrier while](#)
Eddie Griffin crashed a rare Ferrari Enzo worth \$1.5 million into a concrete barrier while
Mar 27, 2007
www.youtube.com

[Ferrari Enzo crash](#)
A real accident of Ferrari Enzo!!! the conductor is Eddie
Mar 27, 2007

Google Search API: References

- Google SOAP Search API
<http://code.google.com/apis/soapsearch/>
- Google AJAX Search API
<http://code.google.com/apis/ajaxsearch/>
- Google AJAX Search API Developer Guide
<http://code.google.com/apis/ajaxsearch/documentation/>
- Google AJAX Search API Samples
<http://code.google.com/apis/ajaxsearch/samples.html>

Lucene



Lucene

Lucene

- Doug Cutting's grandmother's middle name
- Cross-Platform API
- Implemented in Java
 - Ported in C++, C#, Perl, Python
- Offers scalable, high-performance indexing
 - Incremental indexing as fast as batch indexing
 - Index size roughly 20-30% the size of indexed text
- Supports many powerful query types

Lucene: Modules

- Analysis
 - Tokenization, Stop words, Stemming, etc.
- Document
 - Unique ID for each document
 - Title of document, date modified, content, etc.
- Index
 - Provides access and maintains indexes.
- Query Parser
- Search / Search Spans

Lucene: Indexing

- A Document is a collection of Fields



- A Field is free text, keywords, dates, etc.
- A Field can have several characteristics
 - indexed, tokenized, stored, term vectors
 - Apply Analyzer to alter Tokens during indexing
 - Stemming
 - Stop-word removal
 - Phrase identification

Lucene: Searching

- Uses a modified Vector Space Model
- We convert a user's query into an internal representation that can be searched against the Index
- Queries are usually analyzed in the same manner as the index
- Get back `HITS` and use in an application

Lucene: Query Parser Syntax

- Terms
 - Single terms and phrases
- Fields
 - E.g. title:"Do it right" AND right
- Wildcard Searches
 - '?' for single character
 - '*' for multiple characters
- Proximity Searches
 - "jakarta apache"~10
- Fuzzy Searches
 - Levenshtein Distance or Edit Distance algorithm
- Range Searches
 - mod_date:[20020101 TO 20030101]
 - title:{Aida TO Carmen}
- Boosting a Term
 - E.g. jakarta^4 apache
- Boolean Operators

Lucene: More Advanced Options

- Relevance Feedback
 - Manual
 - User selects which documents are relevant/non-relevant
 - Get the terms from the term vector for each of the documents and construct a new query.
 - Automatic
 - Application assumes the top X documents are relevant and the bottom Y are non-relevant and constructs a new query based on the terms in those documents.
 - Span Queries
 - Phrase Matching
-

Lucene: Basic Demo

- The latest version can be obtained from <http://www.apache.org/dyn/closer.cgi/lucene/java/>
- To build an index just type
 - `java org.apache.lucene.demo.IndexFiles <dir>`
- To search from an index type:
 - `java org.apache.lucene.demo.SearchFiles <index>`

Terrier 

Terrier

EPSRC

Engineering and Physical Sciences
Research Council

Terrier: Overview

- Stands for **TER**abyte **RetriE**ve**R**.
- Open Source API (Mozilla Public Licence).
- Written in cross-platform Java.
- Highly compressed disk data structures.
- Handling large-scale document collections.
- Standard evaluation of TREC ad-hoc and known-item search retrieval results.

Terrier: Indexing

- Create your own Collection decoder and Document implementation.
 - Centralized or distributed Setting.
- Indexer iterates through the collection and creates the following data structures
 - Direct Index
 - Document Index
 - Lexicon

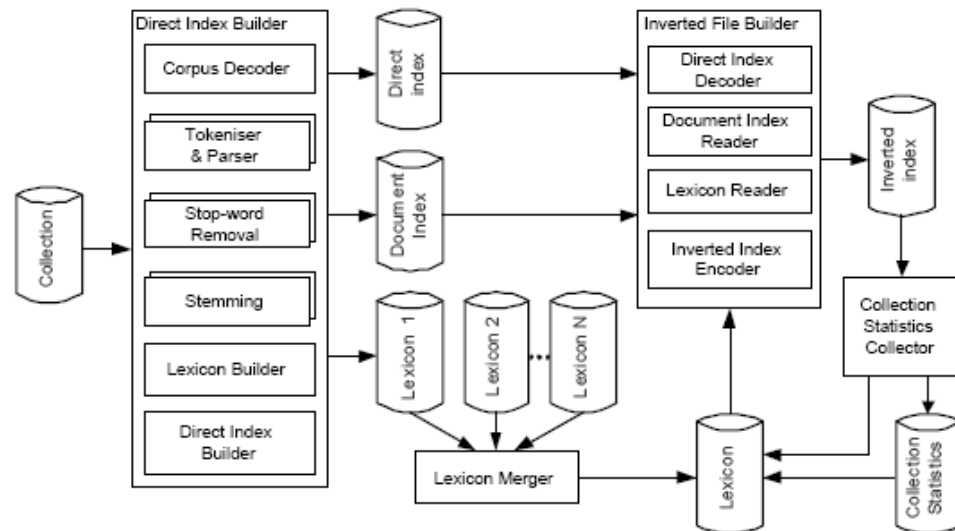
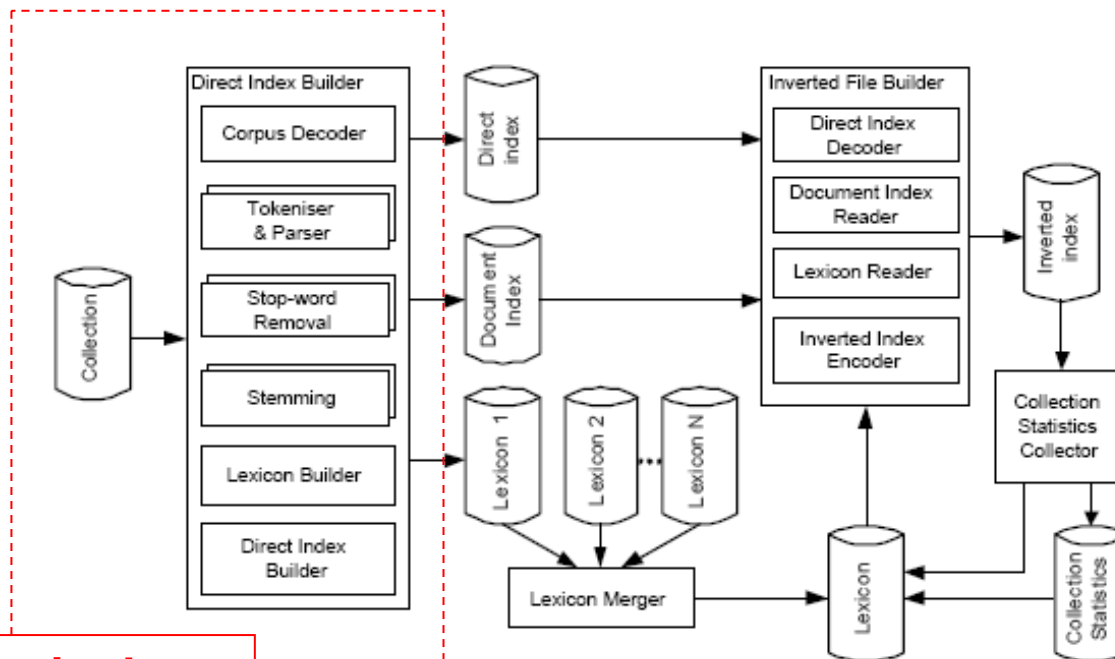


Fig. 1. Indexing process with Terrier.

Terrier: Indexing



Each document in the collection is tokenized and parsed.

Fig. 1. Indexing process with Terrier.

Terrier: Indexing

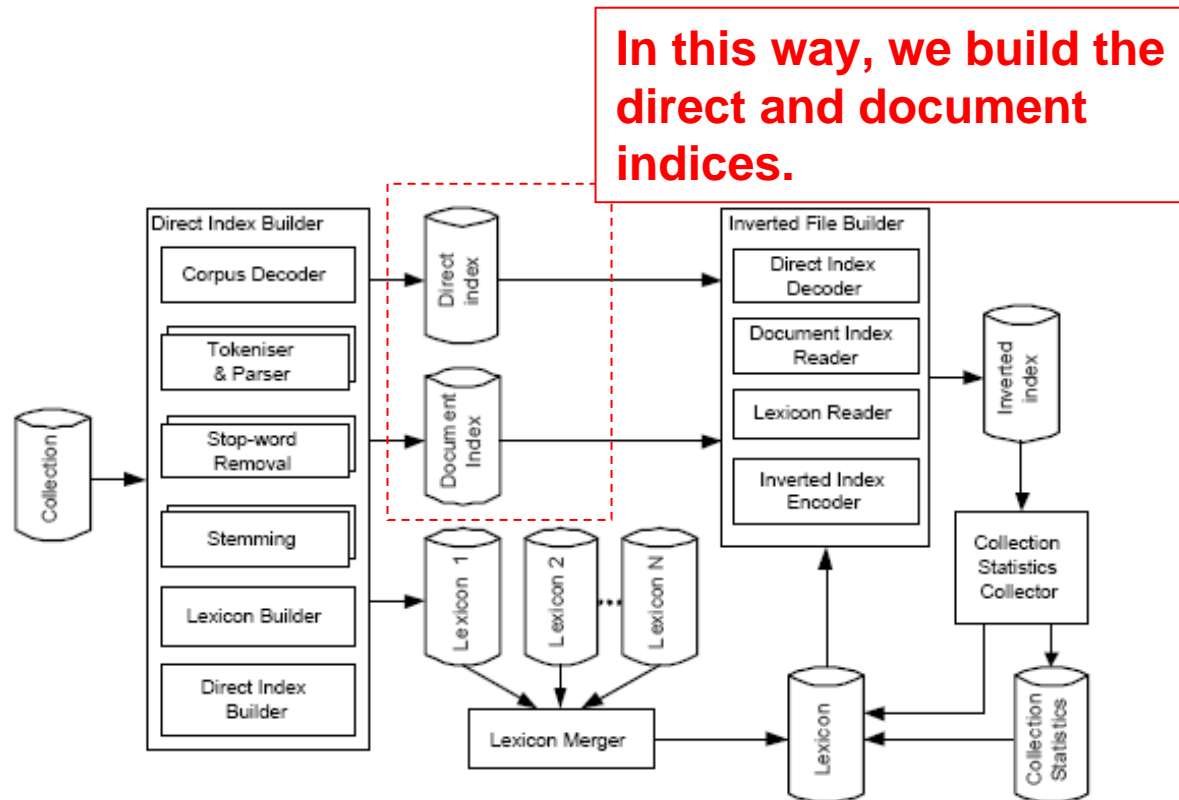


Fig. 1. Indexing process with Terrier.

Terrier: Indexing

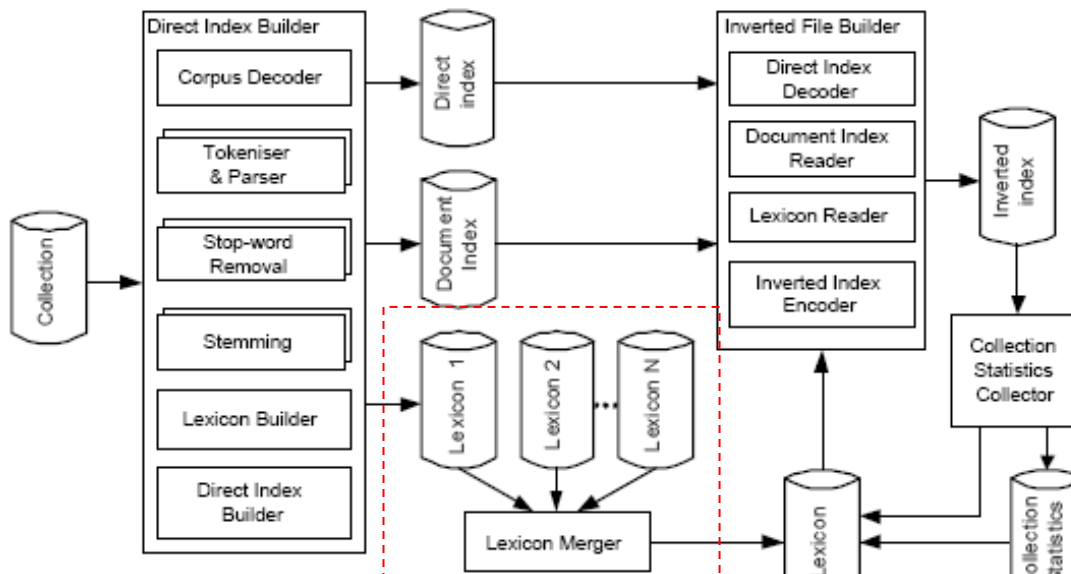


Fig. 1. Indexing process

We also build temporary lexicons in order to reduce the required memory during indexing

Terrier: Indexing

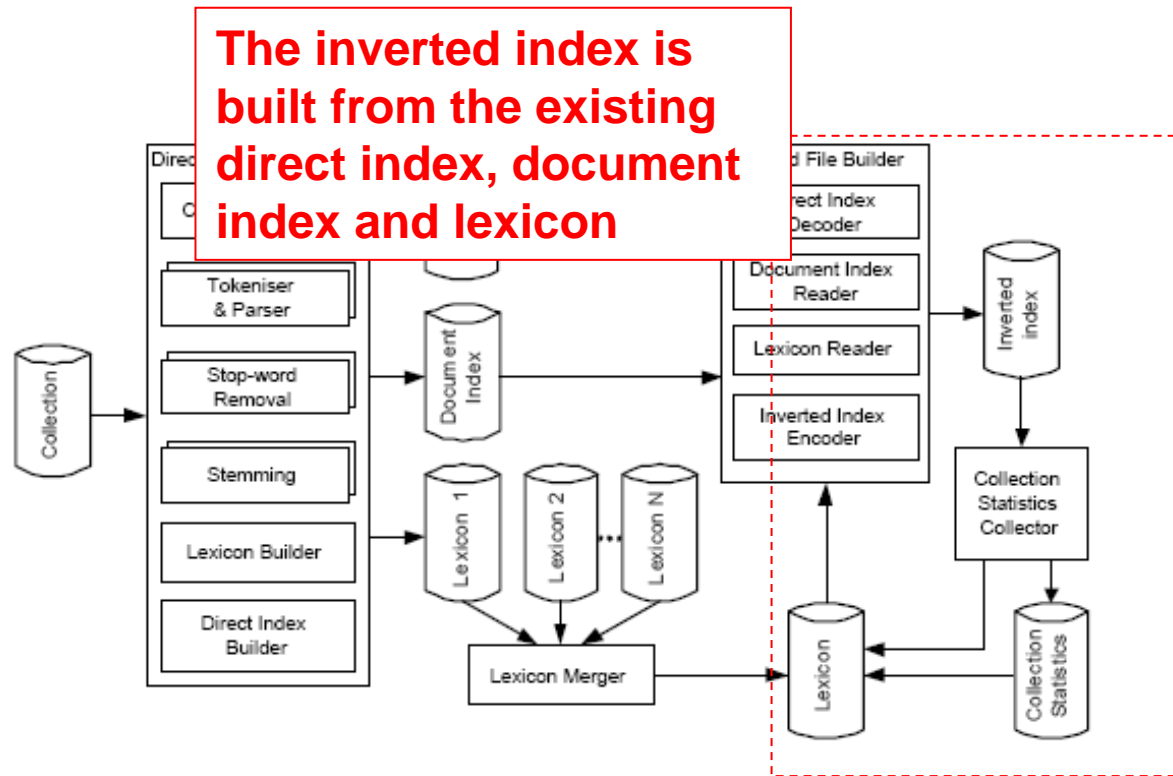


Fig. 1. Indexing process with Terrier.

Terrier: Retrieval

- Parsing
- Pre-processing
- Matching
- Post Processing
- Post Filtering
- Query Language
 - ❑ term1 term2
 - ❑ term1^2.3
 - ❑ +term1 -term2
 - ❑ "term1 term2"~n

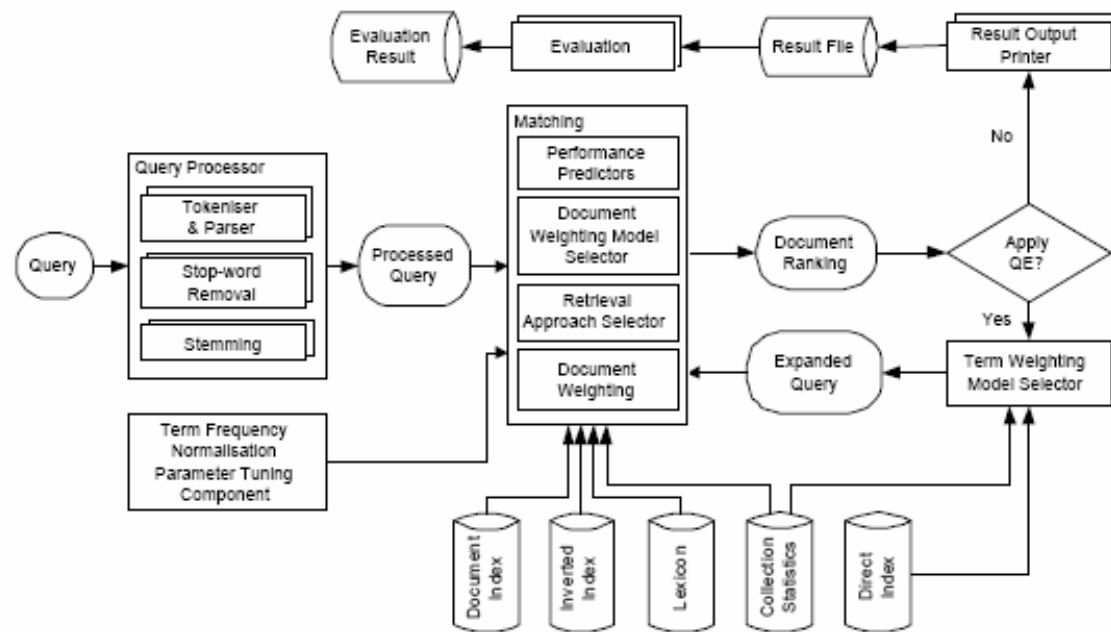
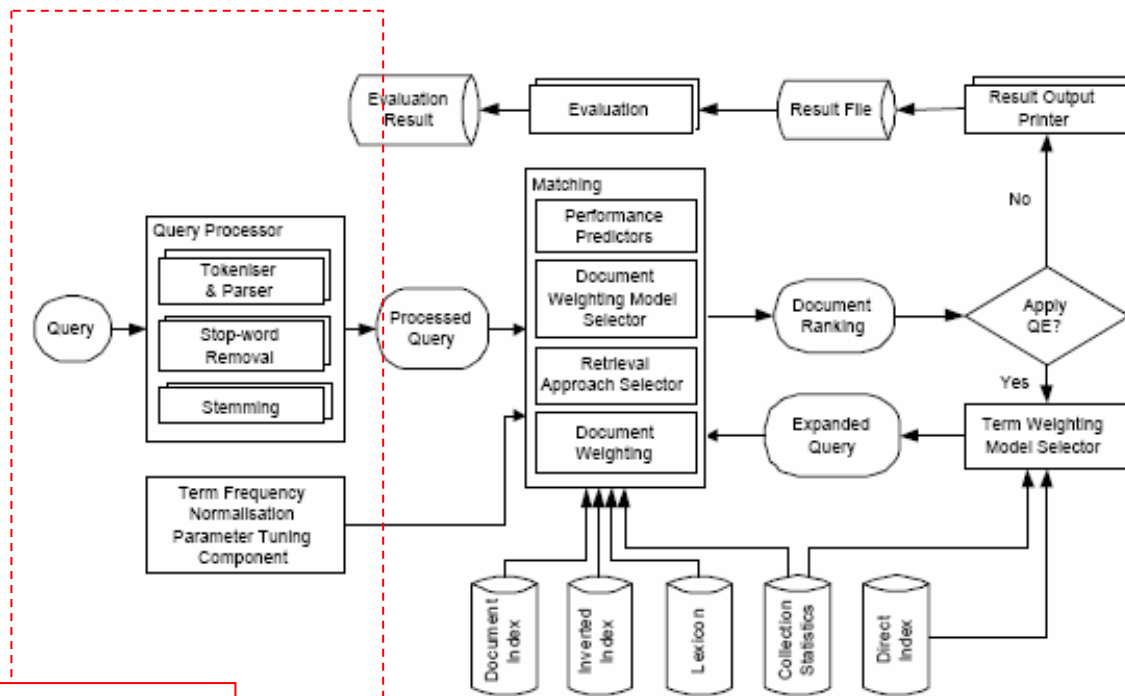


Fig. 2. Retrieval process with Terrier.

Terrier: Retrieval



**Remove stop words
and apply stemming to
the query.**

Fig. 2. Retrieval process with Terrier.

Terrier: Retrieval

Terrier automatically select the optimal document weighting model

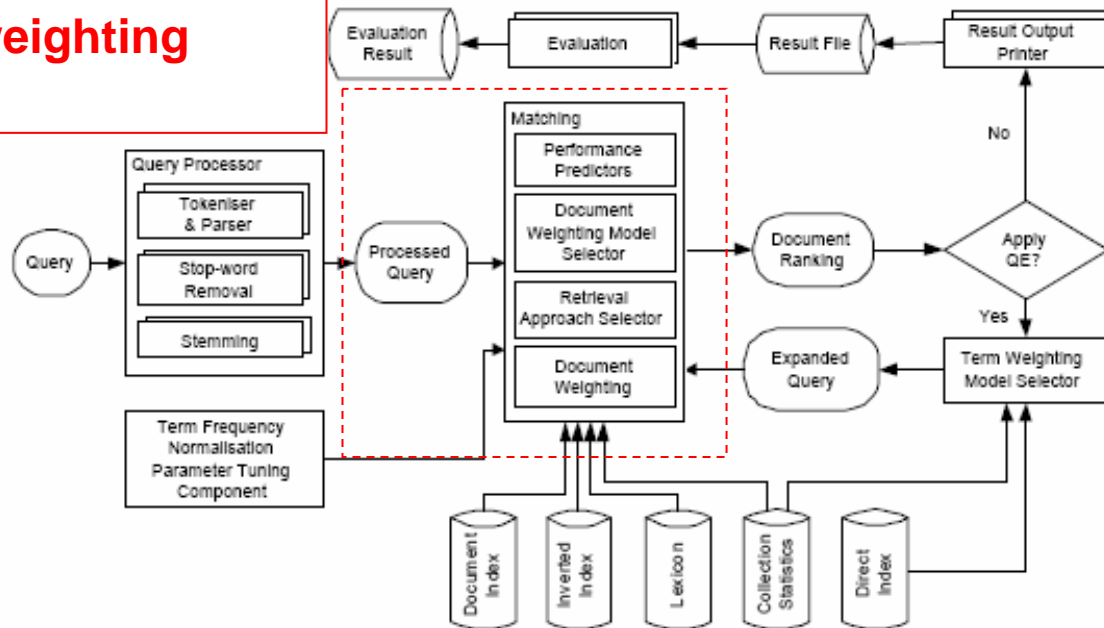


Fig. 2. Retrieval process with Terrier.

Terrier: Retrieval

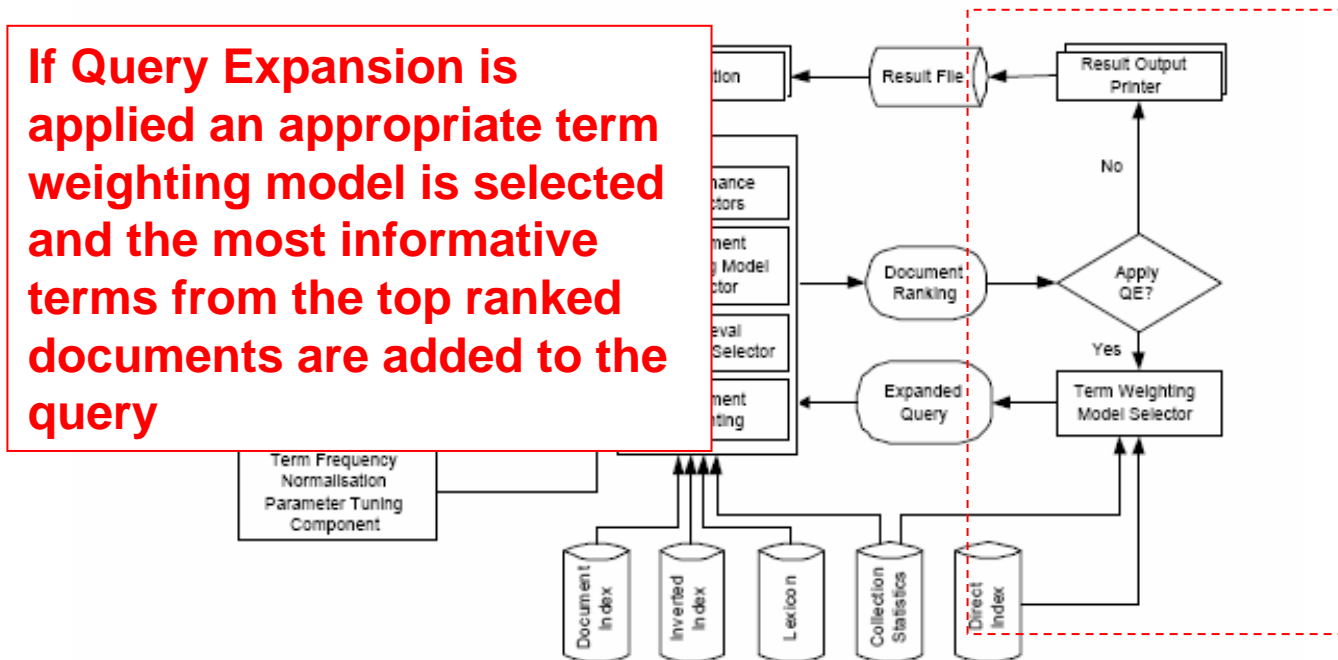


Fig. 2. Retrieval process with Terrier.

Terrier: Sample Applications

- Trec Terrier

- An application that allows Terrier to index and retrieve from standard TREC test collections.
- Instructions are available at http://ir.dcs.gla.ac.uk/terrier/doc/trec_terrier.html

Terrier: Sample Applications

■ Desktop Search

- A Swing (graphical) application that can be used to index files from the local machine, and then perform queries on them.
- The scripts for running the desktop search application are:
 - `desktop_search.sh` (Linux, Mac OSX)
 - `desktop_search.bat` (Windows)

Terrier: Sample Applications

- Interactive Querying
 - A console application for performing simple queries on an existing index and seeing which documents are returned.
 - The scripts for running the console application are:
 - `interactive_terrier.sh` (Linux, Mac OS X)
 - `interactive_terrier.bat` (Windows)

Terrier: Demo

The screenshot shows the Terrier Desktop Search application window. The search term "terrier" is entered in the search box. The results table below shows the following data:

	File Type	Filename	Directory	Score
1	HTML	allclasses-frame.html	C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\allclasses-frame.html	5.7024
2	HTML	allclasses-noframe.html	C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\allclasses-noframe.h...	5.7024
3	HTML	dfr_description.html	C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\dfr_description.html	2.9883
4	HTML	CollectionResultSet.html	C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\matc...	2.5868
5	HTML	TRECFullTokenizer.html	C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\indexi...	2.5474
6	HTML	ResultSet.html	C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\matc...	2.5038
7	HTML	MatchingQueryTerms.html	C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\matc...	2.3853
8	HTML	BitFile.html	C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\comp...	2.3599
9	HTML	Lexicon.html	C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\struct...	2.3381

Below the table, the application displays the following text:

```
NEXT: C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\applications\desktop\p\renaming\class-use\applicationselector.html
NEXT: C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\applications\desktop\filehandling\class-use\AssociationFileOpener.htm
NEXT: C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\applications\desktop\filehandling\class-use\FileOpener.html
NEXT: C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\applications\desktop\filehandling\class-use\MacOSXFileOpener.html
NEXT: C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\applications\desktop\filehandling\class-use\WindowsFileOpener.html
NEXT: C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\matching\models\language\models\class-use\LanguageModel.html
NEXT: C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\matching\models\language\models\class-use\PonteCroft.html
NEXT: C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\matching\models\queryexpansion\class-use\Bo1.html
NEXT: C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\matching\models\queryexpansion\class-use\Bo2.html
NEXT: C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\matching\models\queryexpansion\class-use\KL.html
NEXT: C:\Documents and Settings\gvasil\Επιφάνεια εργασίας\open source\terrier\terrier\doc\javadocl\k\c\g\l\terrier\matching\models\queryexpansion\class-use\QueryExpansionModel.htm
Collection #0 took 1 seconds to block index

flush direct index
flushing block lexicon to disk after the direct index completed
Started building the inverted index...
creating block inverted index
time to process part of lexicon: 0.016
time to traverse direct file: 0.14
time to write inverted file: 0.078
time to perform one iteration: 0.234
number of pointers processed: 35296
Finished building the inverted index...
Time elapsed for inverted file: 0
weighting model: PL2c1.0
1: terrier with 426 documents (TF is 5768).
number of retrieved documents: 426
```




Lemur

Lemur: Overview

- Support for XML and structured document retrieval
- Interactive interfaces for Windows, Linux, and Web
- Cross-Platform, fast and modular code written in C++
- C++, Java and C# APIs
- Free and open-source software

Lemur: API

- Provides interfaces to Lemur classes that are grouped at three different levels:
 - Utility level
 - Common utilities, such as memory management, document parsing, etc.
 - Indexer level
 - Converts a raw text collection to data structures for efficient retrieval.
 - Retrieval level
 - Abstract classes for a general retrieval architecture and concrete classes for several specific information retrieval

Lemur: Indexing

- Multiple indexing methods for small, medium and large-scale (terabyte) collections.
- Built-in support for English, Chinese and Arabic text.
- Porter and Krovetz word stemming.
- Incremental indexing.

Lemur: Retrieval

- Supports major language modelling approaches such as Indri and KL-divergence, as well as vector space, tf-idf, Ocapri and InQuery
- Relevance- and pseudo-relevance feedback
- Wildcard term expansion (using Indri)
- Supports arbitrary document priors (e.g., Page Rank, URL depth)

Questions ?

