# PROJECT REPORT

(Final Year Project 2007-2008)

## Hybrid Search Engine


Project Supervisor

Mrs. Shikha Mehta

# INTRODUCTION

## Search Engines

*Definition:*

A search engine is an information retrieval system designed to help find information stored on a computer system, such as on the World Wide Web, inside a corporate or proprietary network, or in a personal computer.

*Examples:*

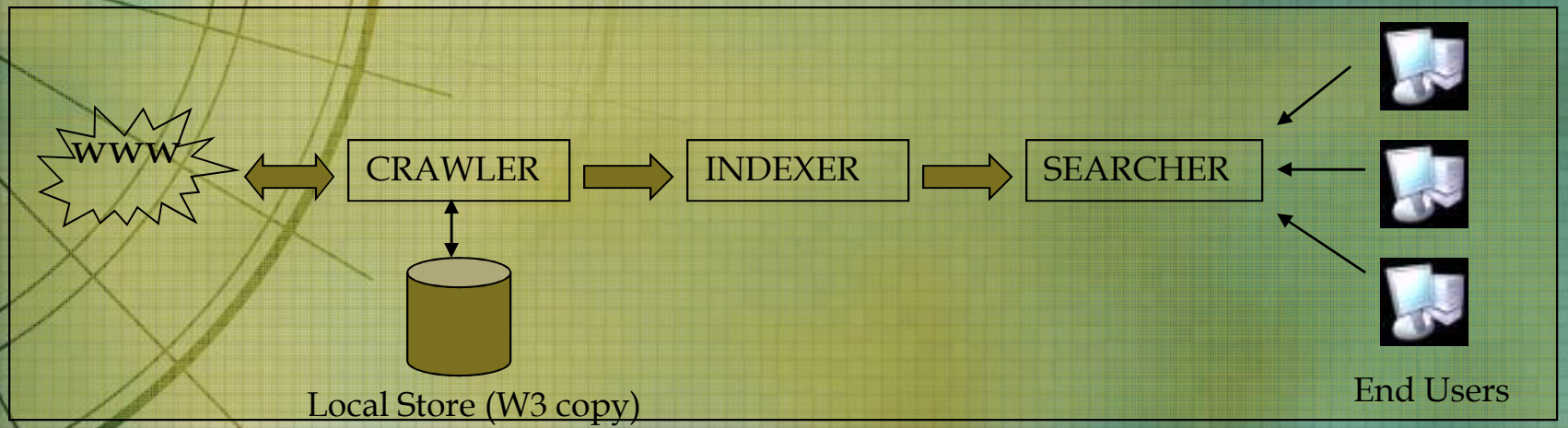
Various search engines are available on the internet e.g. Google, Alta Vista, Ask.com, Yahoo, Lycos, Alltheweb, Myspace, etc.

The popularity of search engines can be estimated by the fact that approximately $112 * 10^6$ searches are made in a single day from one search engine alone.

# How do Search Engines work ?

There are differences in the ways various search engines work, but they all perform three basic tasks:

- They search the Internet -- or select pieces of the web -- based on important words. [CRAWLER]
- They keep an index of the words they find, and where they find them. [INDEXER]
- They allow users to look for words or combinations of words found in that index. [SEARCHER]



# PROBLEM STATEMENT

On the basis of recent studies made on the structure and dynamics of the web itself, it has been analyzed that the web is still growing at a high pace, and the dynamics of the web is shifting. More and more dynamic and real-time information is made available on the web.

Our aim is to design a search engine that meets the challenges of web growth and update dynamics.

# How is our Search Engine "Hybrid" ?

## FAST Crawler

*Features included in our search engine:*

- Freshness algorithm
- Heterogeneous Crawlers
- Heterogeneous Updation mechanism

## COBWeb

*Features included in our search engine:*

- Distributed Architecture
- Inclusion of *"Importance Number"*
- Content based signatures for "Page seen" problem

## HITS

- Link based indexing

## HYBRID SEARCH ENGINE

## PAGE RANK

- Source based indexing focusing on quality of source.

## DOMINOS

*Features included in our search engine:*

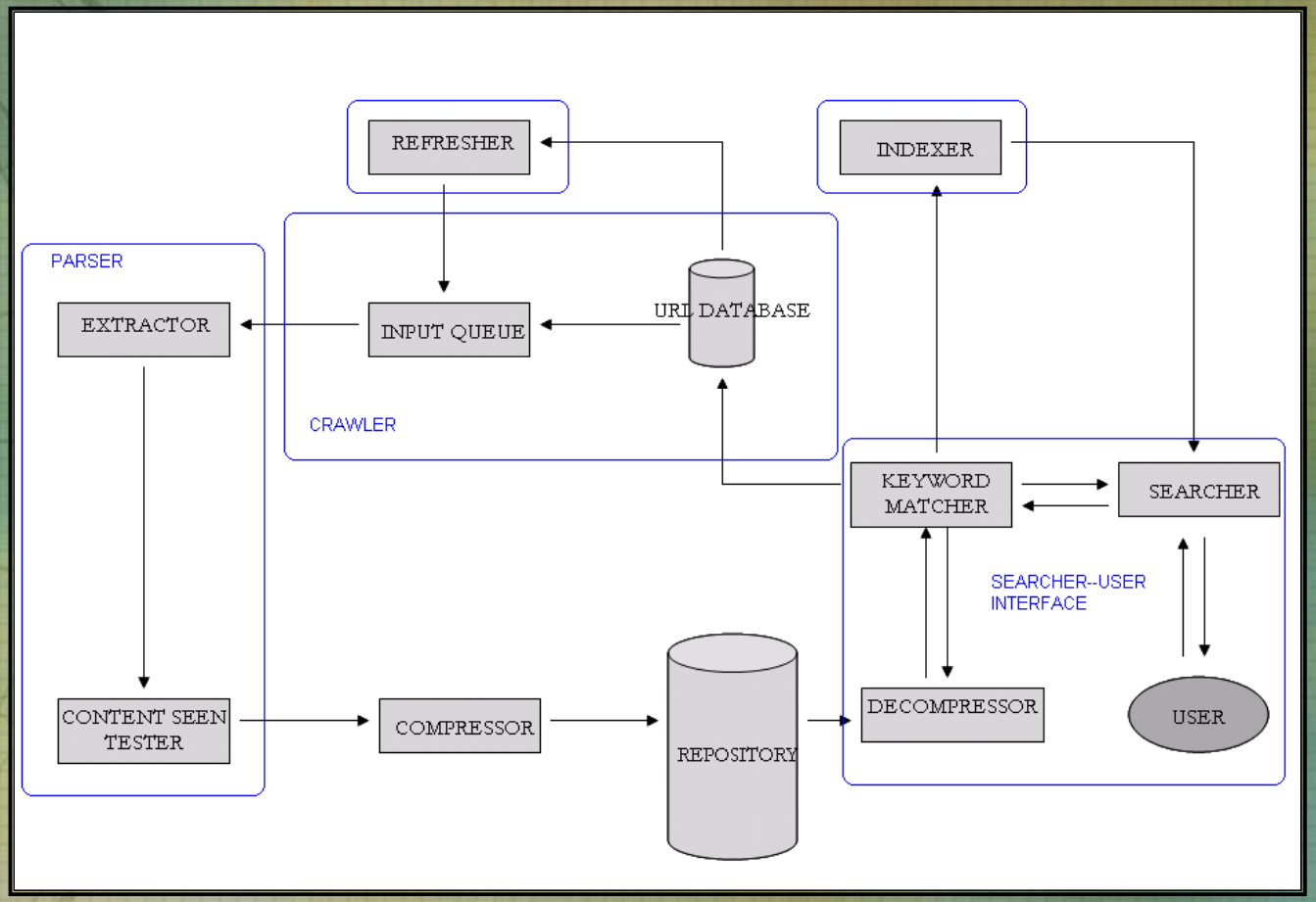
- Inclusion of a new module namely, Local Cache which stores the URLs recently visited.

## Mercator

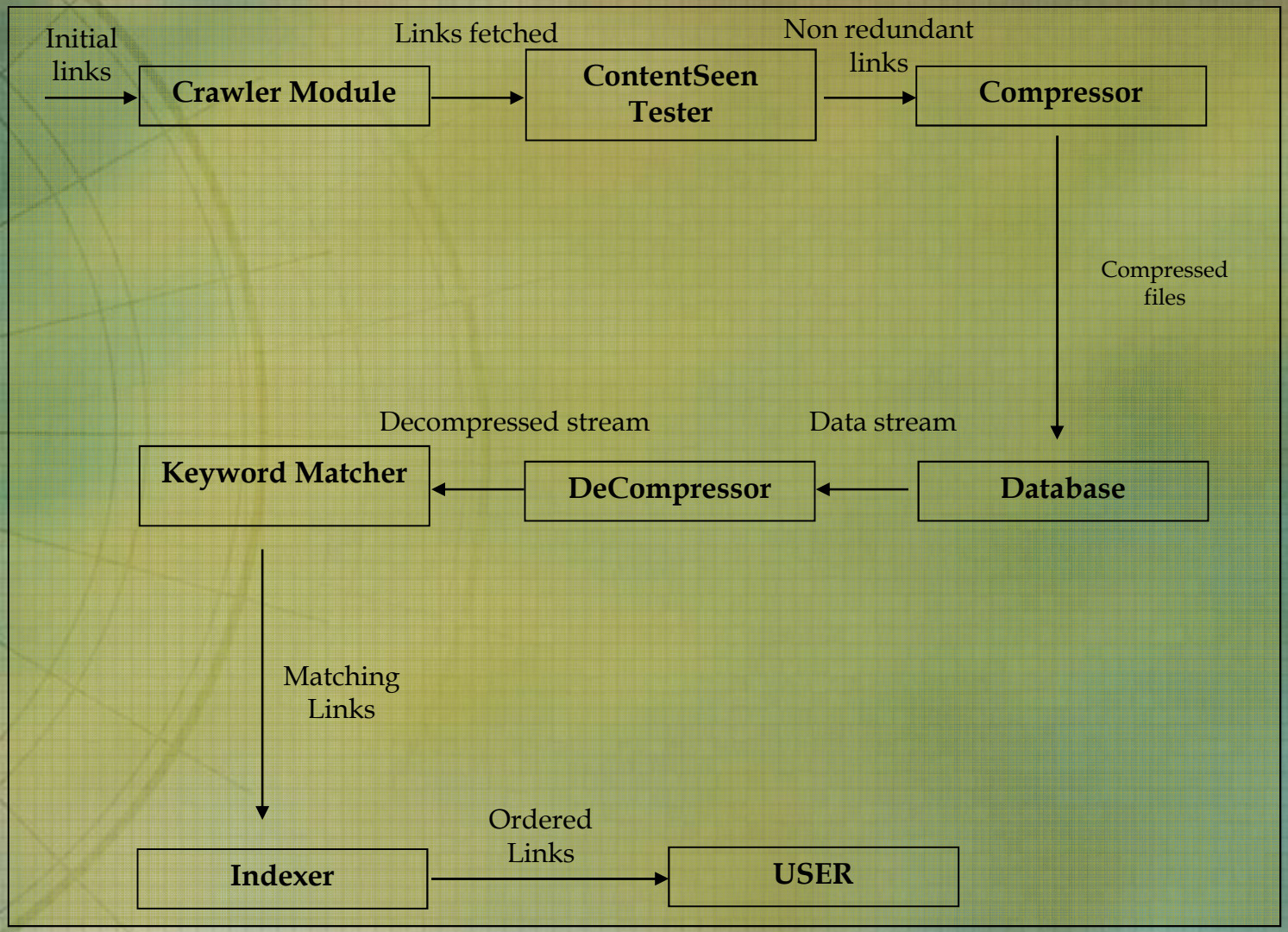
*Features included in our search engine:*

- Content seen test using fingerprinting
- Checkpointing
- URL saving "Checksum" technique

# Our Proposed Design



# Data Flow


Initial links
Crawler Module
Links fetched
ContentSeen Tester
Non redundant links
Compressor
Compressed files
Database
Data stream
DeCompressor
Decompressed stream
Keyword Matcher
Matching Links
Indexer
Ordered Links
USER

# Snapshots

Crawler

**INPUT**

Initial set of URL's taken for the sample :

| Url | Depth |
|---|---|
| http://www.google.com/dirhp/ | 1 |
| http://www.sites-internationaux.com/ | 1 |
| http://www.diroot.com/ | 1 |
| http://www.urlz.net/ | 1 |
| http://www.dmoz.org/ | 1 |
| http://www.hotvsnot.com/ | 1 |
| http://www.webworldindex.com/ | 1 |

## OUTPUT

As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the **crawl frontier**. URLs from the frontier are recursively visited according to a set of policies.

| Url | Depth |
|---|---|
| http://www.webworldindex.com/phtml/USA_Real_Estate/Alabama_Real_Estate/ | 2 |
| http://www.webworldindex.com/phtml/USA_Real_Estate/California_Real_Estate/ | 2 |
| http://www.webworldindex.com/phtml/USA_Real_Estate/Colorado_Real_Estate/ | 2 |
| http://www.webworldindex.com/phtml/USA_Real_Estate/Florida_Real_Estate/ | 2 |
| http://www.webworldindex.com/phtml/USA_Real_Estate/Indiana_Real_Estate/ | 2 |
| http://www.webworldindex.com/phtml/USA_Real_Estate/Maryland_Real_Estate/ | 2 |
| http://www.webworldindex.com/phtml/USA_Real_Estate/North_Carolina_Real_Estate/ | 2 |
| http://www.webworldindex.com/phtml/USA_Real_Estate/South_Carolina_Real_Estate/ | 2 |
| http://www.webworldindex.com/phtml/USA_Real_Estate/Texas_Real_Estate/ | 2 |
| http://www.webworldindex.com/phtml/USA_Real_Estate/Virginia_Real_Estate/ | 2 |
| http://www.webworldindex.com/phtml/Regional/ | 2 |
| http://www.webworldindex.com/phtml/Regional/Countries/ | 2 |

# HTML Parser

**Programming Language**

C#

**Input**

After the crawler crawls the web and store the pages in the repository , we need to extract the useful information from the web page like title , no. of forward links etc. of all the web pages.

**Output**

The extracted information is then stored in the database .

# Output

| Url | Nof | Bit | Uri | Title | Buffer |
|---|---|---|---|---|---|
| http://www.webworldindex.com/phtml/People/Families/ | 36 | a04da0ae4c8a302390bdaabcb7e5cfab | F:\STUDY MATERIAL\Final Year Project\SEARCH ENGINE\Database\test1.txt | Families Directory of Families Websites | 15872 |
| http://www.webworldindex.com/phtml/Recreation_and_Sports/ | 55 | 53692463eb8a6f4a9147d38a20acebbc | F:\STUDY MATERIAL\Final Year Project\SEARCH ENGINE\Database\test2.txt | Recreation and Sports Directory of Recreation and Sports Websites | 24269 |
| http://www.webworldindex.com/phtml/Recreation_and_Sports/Gambling/ | 53 | afd860964c743e2f344297f188542f56 | F:\STUDY MATERIAL\Final Year Project\SEARCH ENGINE\Database\test3.txt | Gambling > Casino Directory > Web World Directory | 21917 |
| http://www.webworldindex.com/phtml/Recreation_and_Sports/Sports/ | 148 | 32a6a1afdcd331cd1e1e763f8600a3b6 | F:\STUDY MATERIAL\Final Year Project\SEARCH ENGINE\Database\test4.txt | Sports Directory of Sports Websites | 41075 |
| http://www.webworldindex.com/phtml/Education/Reference/Libraries/ | 30 | 9cbb66e4a992f6bd7e130d1fe36eac99 | F:\STUDY MATERIAL\Final Year Project\SEARCH ENGINE\Database\test5.txt | Libraries Directory of Libraries Websites | 16487 |
| http://www.webworldindex.com/phtml/Education/Reference/Thesauri/ | 20 | 41c2c65f8ff1e8aa557f535ba07473ca | F:\STUDY MATERIAL\Final Year Project\SEARCH ENGINE\Database\test6.txt | Thesauri Directory of Thesauri Websites | 11742 |
| http://www.webworldindex.com/phtml/Science/Agriculture/ | 61 | 9fdf3183a263cd12e8f729820fd297d4 | F:\STUDY MATERIAL\Final Year Project\SEARCH ENGINE\Database\test7.txt | Agriculture Directory of Agriculture Websites | 21747 |
| http://www.webworldindex.com/phtml/Science/Biology/ | 68 | 3bfab2877017644a6b1ac87b22d25d4e | F:\STUDY MATERIAL\Final Year Project\SEARCH ENGINE\Database\test8.txt | Biology Directory of Biology Websites | 23350 |

Compressor-Decompressor

**Programming Language**

C#

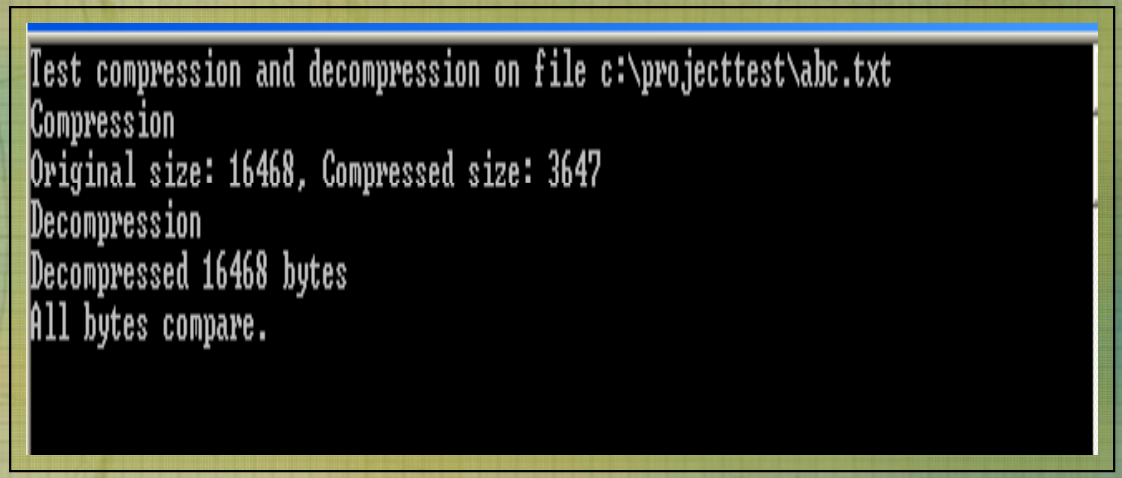
**Input**

After the crawler crawls the web , this module compress the pages and store them in the repository . We need to decompress all the web pages to search a keyword.

**Output**

The compressed pages are stored in the database .

# Output


Test compression and decompression on file c:\projecttest\abc.txt
Compression
Original size: 16468, Compressed size: 3647
Decompression
Decompressed 16468 bytes
All bytes compare.

# Content Seen Tester

**Programming Language**

C#

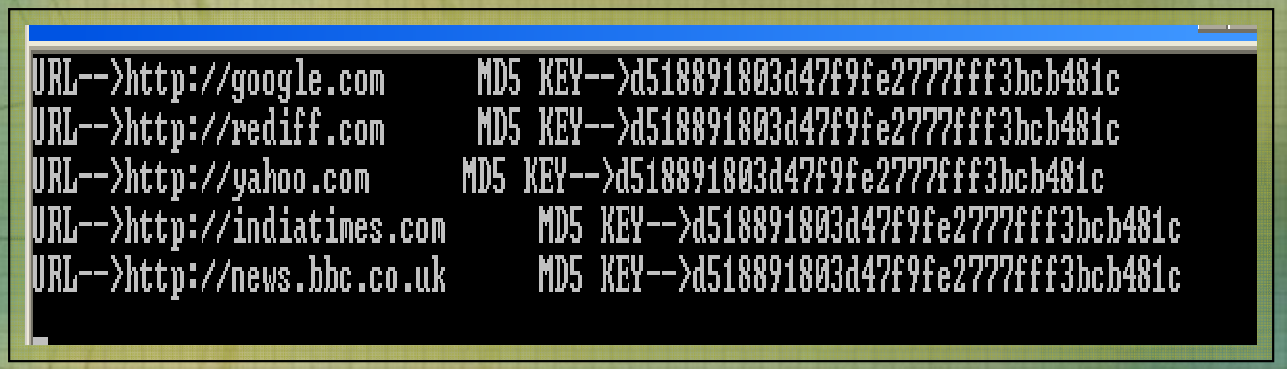
**Input**

The content seen tester generates a bit sequence of all the web pages using MD5 algorithm.

**Output**

The bit sequence of every web page is stored in the database.

# Output


URL-->http://google.com MD5 KEY-->d518891803d47f9fe2777fff3bcb481c
URL-->http://rediff.com MD5 KEY-->d518891803d47f9fe2777fff3bcb481c
URL-->http://yahoo.com MD5 KEY-->d518891803d47f9fe2777fff3bcb481c
URL-->http://indiatimes.com MD5 KEY-->d518891803d47f9fe2777fff3bcb481c
URL-->http://news.bbc.co.uk MD5 KEY-->d518891803d47f9fe2777fff3bcb481c

# Indexer

Sorts the results found on the basis of a rank distribution algorithm.

**Programming Language**

C#

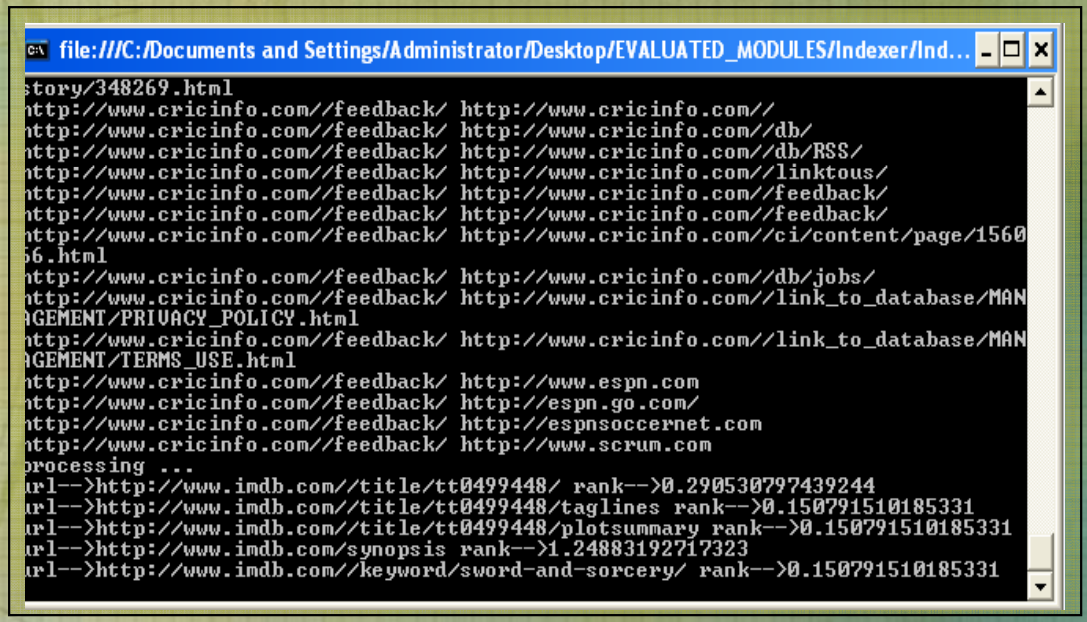
**Input**

The links between all the web pages are fetched from the database.

**Output**

The rank of each web page is stored in the database.

# Output


file:///C:/Documents and Settings/Administrator/Desktop/EVALUATED_MODULES/Indexer/Ind...
story/348269.html
http://www.cricinfo.com//feedback/ http://www.cricinfo.com//
http://www.cricinfo.com//feedback/ http://www.cricinfo.com//db/
http://www.cricinfo.com//feedback/ http://www.cricinfo.com//db/RSS/
http://www.cricinfo.com//feedback/ http://www.cricinfo.com//linktous/
http://www.cricinfo.com//feedback/ http://www.cricinfo.com//feedback/
http://www.cricinfo.com//feedback/ http://www.cricinfo.com//feedback/
http://www.cricinfo.com//feedback/ http://www.cricinfo.com//ci/content/page/1560
56.html
http://www.cricinfo.com//feedback/ http://www.cricinfo.com//db/jobs/
http://www.cricinfo.com//feedback/ http://www.cricinfo.com//link_to_database/MAN
AGEMENT/PRIVACY_POLICY.html
http://www.cricinfo.com//feedback/ http://www.cricinfo.com//link_to_database/MAN
AGEMENT/TERMS_USE.html
http://www.cricinfo.com//feedback/ http://www.espn.com
http://www.cricinfo.com//feedback/ http://espn.go.com/
http://www.cricinfo.com//feedback/ http://espnsoccernet.com
http://www.cricinfo.com//feedback/ http://www.scrum.com
processing ...
url-->http://www.imdb.com//title/tt0499448/ rank-->0.290530797439244
url-->http://www.imdb.com//title/tt0499448/taglines rank-->0.150791510185331
url-->http://www.imdb.com//title/tt0499448/plotsummary rank-->0.150791510185331
url-->http://www.imdb.com/synopsis rank-->1.24883192717323
url-->http://www.imdb.com//keyword/sword-and-sorcery/ rank-->0.150791510185331

# Refresher

Updates the local database with fresh copies of web pages.

**Programming Language**

C#

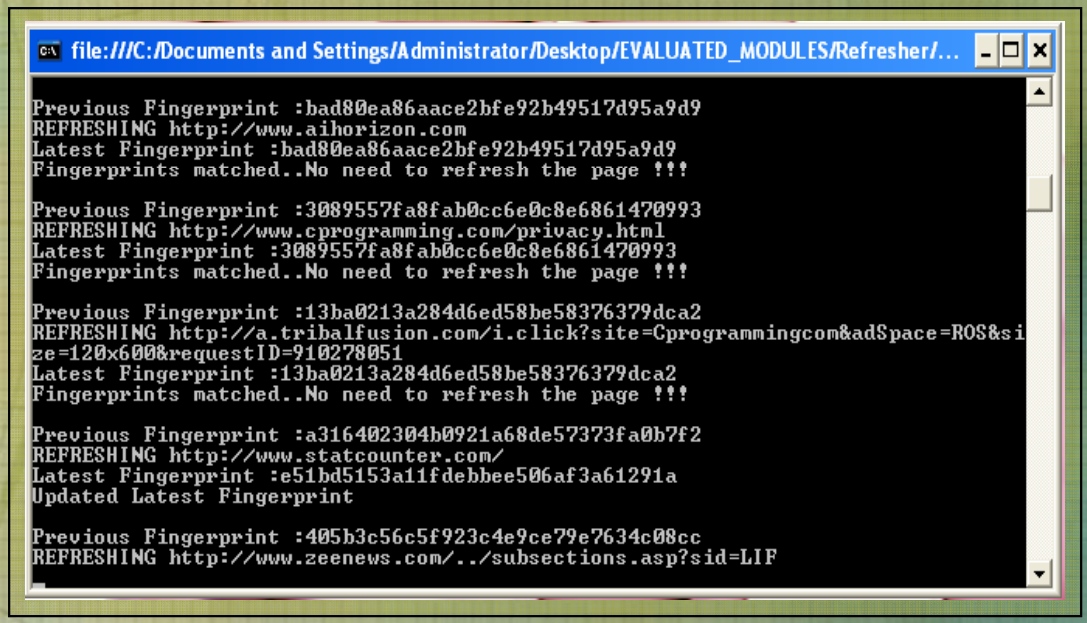
**Input**

The cached pages from the database.

**Output**

The refreshed pages are stored in the repository.

# Output


file:///C:/Documents and Settings/Administrator/Desktop/EVALUATED_MODULES/Refresher/...
Previous Fingerprint :bad80ea86aace2bfe92b49517d95a9d9
REFRESHING http://www.aihorizon.com
Latest Fingerprint :bad80ea86aace2bfe92b49517d95a9d9
Fingerprints matched..No need to refresh the page !!!
Previous Fingerprint :3089557fa8fab0cc6e0c8e6861470993
REFRESHING http://www.cprogramming.com/privacy.html
Latest Fingerprint :3089557fa8fab0cc6e0c8e6861470993
Fingerprints matched..No need to refresh the page !!!
Previous Fingerprint :13ba0213a284d6ed58be58376379dca2
REFRESHING http://a.tribalfusion.com/i.click?site=Cprogrammingcom&adSpace=ROS&size=120x600&requestID=91027805
Latest Fingerprint :13ba0213a284d6ed58be58376379dca2
Fingerprints matched..No need to refresh the page !!!
Previous Fingerprint :a316402304b0921a68de57373fa0b7f2
REFRESHING http://www.statcounter.com/
Latest Fingerprint :e51bd5153a11fdebbee506af3a61291a
Updated Latest Fingerprint
Previous Fingerprint :405b3c56c5f923c4e9ce79e7634c08cc
REFRESHING http://www.zeenews.com/../subsections.asp?sid=LIF

# User Interface

**Programming Language**

ASP .NET

**Input**

The user enters a keyword or multiple keywords.

**Output**

The results are fetched to the user.

# Output


SWIFT
programming
SEARCH
C++ Programming/Computer Science MegaQuiz - Cprogramming.com
http://www.cprogramming.com/cgi-bin/quiz.cgi
USER RATING :
Visual C++ Compiler - Cprogramming.com
http://www.cprogramming.com/visual.html
USER RATING :
Sun Microsystems Communities
http://www.sun.com//communities/
USER RATING :
C programming.com - Your Resource for C and C++ Programming
http://cprogramming.com/tutorial/sdl/setup.html
USER RATING :
Function Pointers in C and C++ - Cprogramming.com
Done

# Thank You!!

Presented By:

ANKUSH GULATI
040303
Project Id: B103

ANKIT KALRA
040321
Project Id: B119