Our MP3 Search Engine

- Crawler
  - Searching for Artist Name
  - Searching for Song Title
- Website
- Difficulties
- Looking Back

Crawler

- Starts given a list of seeds
- Uses a Priority Queue
  - Associate a priority with a given link
    - Priority depends on keyword
    - Store text from two areas:
      - Surrounding text before and up to the link
      - From the anchor
- Grabs the robot.txt file
  - Keep a cache of 10 most recent

Searching for Artist Name

- Check for artist name in anchor text first, then the text before the link
- Use the UBL.COM site
  - Make sure to be polite
- Three types of matches from UBL
  - No match
  - Many matches
  - Exact match

Artist Name Search Algorithm

- Start with single name searches
  - Ex: Madonna
  - If find exact match assume as name
  - No exact match keep record of many matches
- Move onto two word names
  - Ex: Michael Jackson
  - If find exact match assume as name
  - No exact match keep record of many matches

Artist Name Search Algorithm

  - If no exact match for single or double names
    - Take a guess: If there is a guess recorded from double assume as name, if not take single name guess as the name
- Save the new found artist name into our database

Searching for Song Title

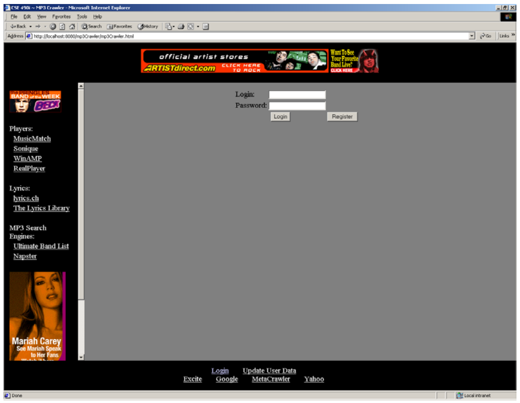
- We only search for song title if we have the artist name
  - Too many song titles
- Wrapper for Audiogalaxy.com
- Given the artist name retrieve all songs, under that artist

Song Title Search Algorithm

- Now we have list of all songs by that artist in our database
  - First search for the song title in the anchor text
  - If not found, then search for song title in the text before and up to the link
  - Once title found store the MP3 link, artist name and song title into our sing_by table in our database

Ranking

- Use the artist name search algorithm for ranking
  - Rankings highest to lowest:
    - Exact match found in anchor (4)
    - Exact match found in surrounding text (3)
    - Guess made from anchor (2)
    - Guess made from surrounding text (1)
    - No artist name was found (0)


CSE 490i ~ MP3 Crawler - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://localhost:8080/mp3Crawler/mp3Crawler.html
Go Links
official artist stores
ARTISTdirect.com CLICK HERE TO ROCK
Want To See Your Favorite Band Live? CLICK HERE
BAND of the WEEK BECK
Players:
MusicMatch
Sonique
WinAMP
RealPlayer
Lyrics:
lyrics.ch
The Lyrics Library
MP3 Search Engines:
Ultimate Band List
Napster
Mariah Carey
See Mariah Speak to Her Fans
Login
Update User Data
Excite
Google
MetaCrawler
Yahoo
Done
Local intranet

Login:

Password:

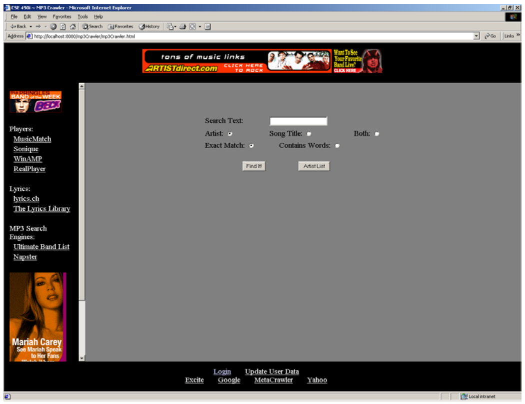
Login Register

Website Features

- Login Page
  - So if a registered user is not on his usual computer
  - If cookie on computer automatically enters into our search page
- Update User Info Page
  - Allow the user to change their info and password

Website Features

- Search Page
  - Search by artist or title
  - Search as exact or "contains"
  - Index of artist names
    - Ex: Click on "A" return artist that start with A

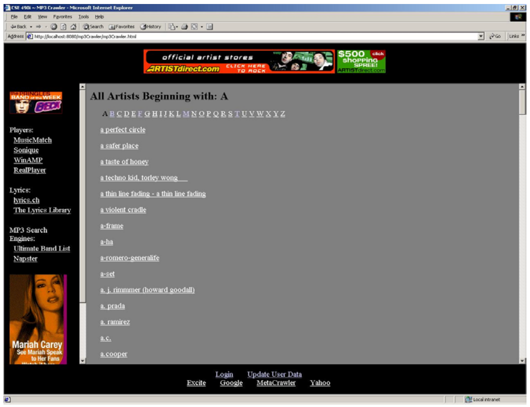

CSE 490i ~ MP3 Crawler - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://localhost:8080/mp3Crawler/mp3Crawler.html
Go Links
tons of music links
ARTISTdirect.com CLICK HERE TO ROCK
Want To See Your Favorite Band Live? CLICK HERE
BAND of the WEEK BECK
Players:
MusicMatch
Sonique
WinAMP
RealPlayer
Lyrics:
lyrics.ch
The Lyrics Library
MP3 Search Engines:
Ultimate Band List
Napster
Mariah Carey
See Mariah Speak to Her Fans
Login
Update User Data
Excite
Google
MetaCrawler
Yahoo
Local intranet

Search Text:

Artist: ● Song Title: ○ Both: ○

Exact Match: ● Contains Words: ○

Find It! Artist List



CSE 490i ~ MP3 Crawler - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://localhost:8080/mp3Crawler/mp3Crawler.html
Go Links
official artist stores ARTISTdirect.com CLICK HERE TO ROCK
$500 shopping SPREE! click ARTISTdirect.com
the PRINGLES BAND of the WEEK BECK
Mariah Carey See Mariah Speak to Her Fans
Local intranet

Players:
MusicMatch
Sonique
WinAMP
RealPlayer

Lyrics:
lyrics.ch
The Lyrics Library

MP3 Search Engines:
Ultimate Band List
Napster

All Artists Beginning with: A

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

a perfect circle

a safer place

a taste of honey

a techno kid, torley wong

a thin line fading - a thin line fading

a violent cradle

a-frame

a-ha

a-romero-generalife

a-set

a. j. rimmmer (howard goodall)

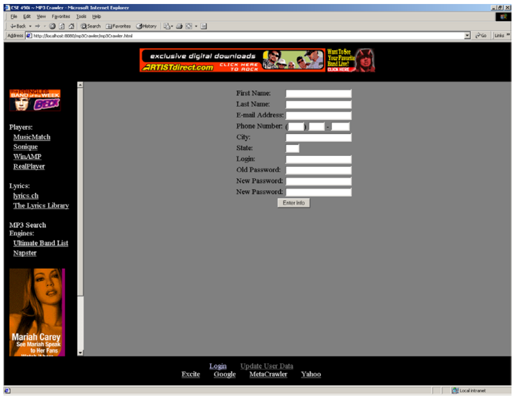
a. prada

a. ramirez

a.c.

a.cooper

Login Update User Data
Excite Google MetaCrawler Yahoo



CSE 490i ~ MP3 Crawler - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://localhost:8080/mp3Crawler/mp3Crawler.html
Go Links
exclusive digital downloads ARTISTdirect.com CLICK HERE TO ROCK
Want To See Your Favorite Band Live? CLICK HERE
the PRINGLES BAND of the WEEK BECK
Mariah Carey See Mariah Speak to Her Fans
Enter Info
Local intranet

Players:
MusicMatch
Sonique
WinAMP
RealPlayer

Lyrics:
lyrics.ch
The Lyrics Library

MP3 Search Engines:
Ultimate Band List
Napster

First Name:
Last Name:
E-mail Address:
Phone Number: ( ) -
City:
State:
Login:
Old Password:
New Password:
New Password:

Login Update User Data
Excite Google MetaCrawler Yahoo

Difficulties

- Initially it was very hard to be polite
  - We repeatedly requested for robot.txt file if the host did not have one
- Our crawler was very slow
  - Searching for artist name and song title were very slow due to politeness policies

More Difficulties

- Running out of Virtual Memory
  - Stored all queues in our database:
    - Links to visit
    - Of Mp3 links to search for artist name
    - Of Mp3 links to search for song title

Looking Back

- Things we learned
  - Crawler issues:
    - What site to visit next
    - Politeness issues
  - Java Servlet & JavaScript
  - Teamwork
  - Make accurate and descriptive documentation (write-ups)
- Next time better designing and planning ahead of time