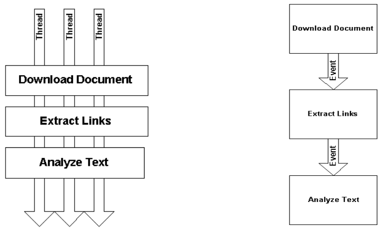
CrawlBuddy

The web's best friend

Design Decisions

- Two paradigms of design to choose from: multi-threaded or event-driven



Multi-Threaded Programming

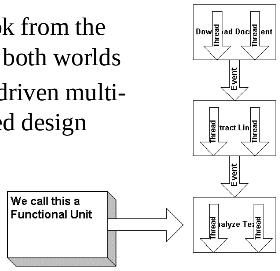
- **Advantages**
- Programming is easier because threads are linear and we (usually) think linearly
- Threads can take advantage of multiprocessors easily
- Threads are synchronous i.e. it is okay for a thread to block because there are many of them running at once
- Debugging a threaded program is considerably easier than an event based program
- **Disadvantages**
- Threads are limited by the underlying operating system (operating systems can only efficiently handle so many threads)

Event-Driven Programming

- **Advantages**
- Handles well under heavy load, the queues act as a buffer to soften the load
- Simple to add new functionality and process in parallel
- Easy to split up and run on multiple machines
- Modular
- **Disadvantages**
- Not as intuitive as Thread programming
- Harder to debug system level errors (but easier to debug individual pieces)

What CrawlBuddy Does

- We took from the best of both worlds
- Event-driven multi-threaded design

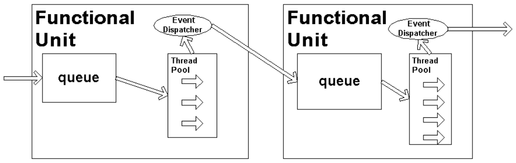


Functional Units Are Our Friends

- Each Functional Unit has a …
- Queue – holds events to be processed
- Thread Pool – takes events off the queue and processes them
- Event Dispatcher – sends events to other Functional Units

Design of a Functional Unit

- Arrows represent flow of a task



Functional Unit

Event Dispatcher

queue

Thread Pool

Functional Unit

Event Dispatcher

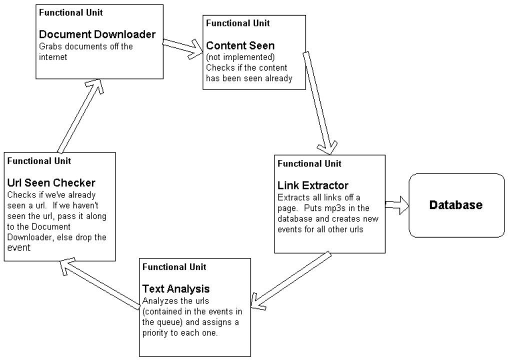
queue

Thread Pool

CrawlBuddy Design

- Basically, events are passed between Functional Units
- The arrows (on the next slide) represent event flow

CrawlBuddy Design Flow



Functional Unit

Document Downloader
Grabs documents off the internet

Functional Unit

Content Seen
(not implemented)
Checks if the content has been seen already

Functional Unit

Link Extractor
Extracts all links off a page. Puts mp3s in the database and creates new events for all other urls

Database

Functional Unit

Text Analysis
Analyzes the urls (contained in the events in the queue) and assigns a priority to each one.

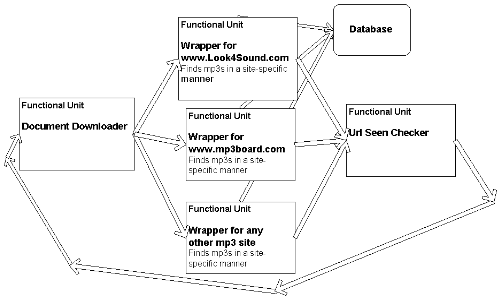
Functional Unit

Url Seen Checker
Checks if we've already seen a url. If we haven't seen the url, pass it along to the Document Downloader, else drop the event

Wrapper Design

- Our wrapper crawler targets specific sites and uses site-specific format to find mp3s and record information about them (song name, artist name, etc)
- The wrapper Functional Units can be run in parallel and the each use the same database
- The Document Downloader passes each event to each of the wrappers. If the event does not apply to the wrapper (i.e. the document comes from a different site), the wrapper will simply drop the event

Wrapper Design Flow



Functional Unit

Document Downloader

Functional Unit

Wrapper for www.Look4Sound.com
Finds mp3s in a site-specific manner

Functional Unit

Wrapper for www.mp3board.com
Finds mp3s in a site-specific manner

Functional Unit

Wrapper for any other mp3 site
Finds mp3s in a site-specific manner

Database

Functional Unit

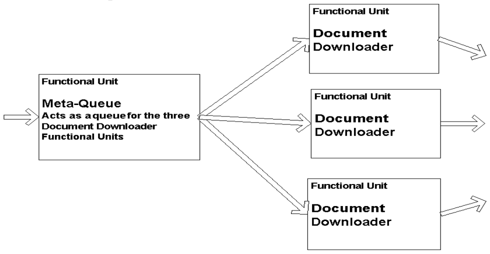
Url Seen Checker

Design Advantages

- Code re-use (Functional Units shared across CrawlBuddy and the wrapper)
- Expandable
- Checkpointing is simple (save the queues)
- Easy to run on multiple machines
- Queues buffer the load on threads
- Functional Units Replicable (see next slide)

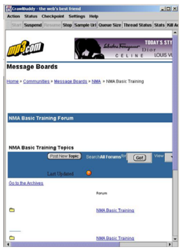
Meta Queue

- How to replicate Functional Units



CrawlBuddy Features

- GUI



Message Boards

Home > Communities > Message Boards > NMA > NMA Basic Training

NMA Basic Training Forum

NMA Basic Training Topics

Post New Topic Search All Forums for Go! View

Last Updated

Go to the Archives

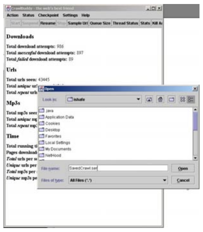
Forum

NMA Basic Training

NMA Basic Training

CrawlBuddy Features (cont)

- Checkpointing



Downloads

Total download attempts: 986
Total *successful* download attempts: 897
Total *failed* download attempts: 89

Urls

Total urls seen: 43445
Total *unique* url
Total *repeat* urls

Mp3s

Total mp3s seen
Total *unique* mp
Total *repeat* mp3

Time

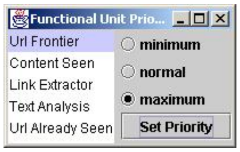
Total running ti
Pages download
***Total* urls per s**
***Unique* urls per**
***Total* mp3s per**
***Unique* mp3s pe**

.java
Application Data
Cookies
Desktop
Favorites
Local Settings
My Documents
NetHood

File name: SavedCrawl.ser
Files of type: All Files (*.*)

CrawlBuddy Features (cont)

- Dynamic control of Functional Unit priority

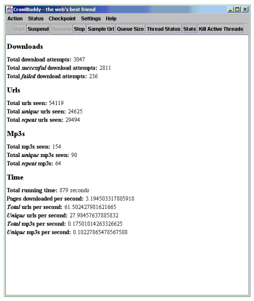


Url Frontier
Content Seen
Link Extractor
Text Analysis
Url Already Seen

○ minimum
○ normal
● maximum

CrawlBuddy Features (cont)

- Real-time stats
- Total downloads
- Total mp3
- Downloads / sec
- Etc.



Downloads

Total download attempts: 3047
Total *successful* download attempts: 2811
Total *failed* download attempts: 236

Urls

Total urls seen: 54119
Total *unique* urls seen: 24625
Total *repeat* urls seen: 29494

Mp3s

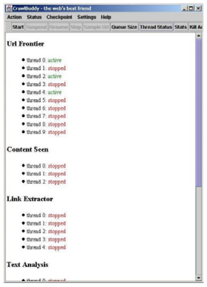
Total mp3s seen: 154
Total *unique* mp3s seen: 90
Total *repeat* mp3s: 64

Time

Total running time: 879 seconds
Pages downloaded per second: 3.194503317805918
***Total* urls per second:** 61.502427981621665
***Unique* urls per second:** 27.98457637885832
***Total* mp3s per second:** 0.17501014263326625
***Unique* mp3s per second:** 0.10227865478567508

CrawlBuddy Features (cont)

- Thread status monitor



Url Frontier

- thread 0: active
- thread 1: stopped
- thread 2: active
- thread 3: stopped
- thread 4: active
- thread 5: stopped
- thread 6: stopped
- thread 7: stopped
- thread 8: stopped
- thread 9: stopped

Content Seen

- thread 0: stopped
- thread 1: stopped
- thread 2: stopped

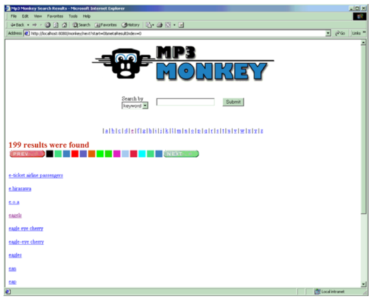
Link Extractor

- thread 0: stopped
- thread 1: stopped
- thread 2: stopped
- thread 3: stopped
- thread 4: stopped

Text Analysis

Mp3Monkey

- Search for all ‘e’ artists


Mp3 Monkey Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address
Go
Links
MP3 MONKEY
Submit
PREV
NEXT
Local intranet

http://localhost:8080/monkey/next?start=0&metaResultIndex=0

Search by keyword

| 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

199 results were found

e-ticket airline passengers

e hirasawa

e.o.a

eagels

eagle eye cherry

eagle-eye cherry

eagles

ean

eap

Mp3Monkey Features

- Self-maintaining database – if a user attempts to download non-existent mp3, that url is marked for deletion
- Statistics are kept of how many searches and what has been downloaded