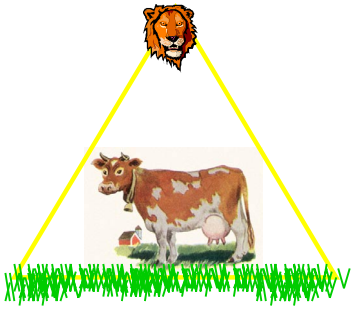
Information Carnivores



Outline

- **MetaCrawler (briefly)**
- **Wrappers**
- **Meta-bots**
  - Jango
- **The Story of Netbot**

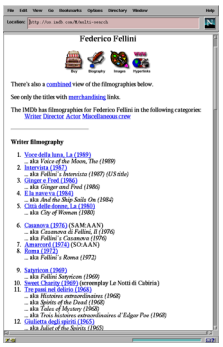
- **Built by Selberg & Etzioni**
- **Release in Jun '95**
- **Currently aggregates 12 search engines:**
  - LookSmart, About, Infoseek,
  - GoTo, Google, DirectHit,
  - RealNames, Webcrawler, AltaVista,
  - Excite, Lycos, Thunderstone

The Need for Wrappers

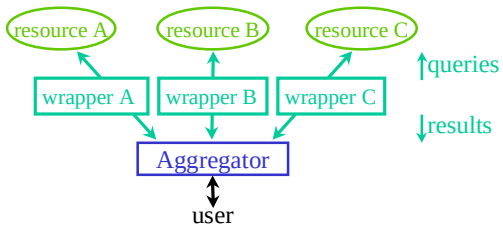
lots of
information

but

computers don't
understand
much of it



Strategy: ***Wrappers***



Scaling issues

Need custom wrapper for *each* resource.

```
<HTML><BODY BGCOLOR="FFFFFF" LINK=
"00009C" ALINK="00009C" VLINK="00009C"
TEXT= "000000"> <center> <table><tr><td><NO
BR> <NOBR><img src="/ypimages/b_r_hd_a.gif"
border=0 ALT="Switchboard Results" width=407
height=20 align=top><A HREF="/bin/cgiqa.dll?MEM
=1" TARGET ="_top"><img src="/ypimages/b_r_hd
_1.gif" border=0 ALT="People" width=54 height=20
align=top></A><A HREF="/bin/cgidir.dll?MEM=1"
TARGET="_top"><img src= "/ypimages/b_r_hd_2.gif"
border=0 ALT= "Business" width=62 height=24 align
=top></A><A HREF="/" TARGET="_top"><img src
="/ypimages /b_r_hd_3.gif" border=0 ALT="Home"
width=47 height=20 align=top></A></NOBR><br>
</td></tr></table> </center><center><table border=0
width=576> <tr><td colspan=2 align =center> <center>
```

→ useful information

But hand-coding is *tedious*.

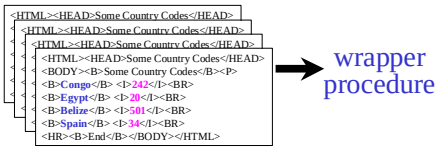
Especially since sites frequently change format

Wrapper Approaches

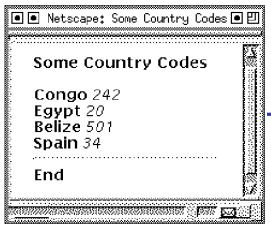
- **Perl-like languages**
  - Simple and effective (if tedious)
- **Proprietary languages & tools**
  - Click and generalize
- **Conversion to tree form**
  - Use XML as intermediate representation
  - Extract children of specified node
- **Machine Learning**
  - Promising, but not yet fielded

Kushmerick Contribution

machine learning techniques to automatically construct wrappers from examples



Example



⟨ (Congo, 242)
(Egypt, 20)
(Belize, 501)
(Spain, 34) ⟩

LR wrappers: The basic idea

exploit ***fortuitous non-linguistic regularity***

<HTML><TITLE>Some Country Codes</TITLE>
<B>Congo</B> <I>242</I>
<B>Egypt</B> <I>20</I>
<B>Belize</B> <I>501</I>
<B>Spain</B> <I>34</I>
</BODY></HTML>

Use **<B>**, **</B>**, **<I>**, **</I>** for parsing

Country/Code LR wrapper

procedure ExtractCountryCodes
while there are more occurrences of <B>
1. **extract Country between <B> and </B>**
2. **extract Code between <I> and </I>**

***L*eft-*R*ight wrappers**

Value of Meta-Information

- **TV guide makes more than NBC**
- **American Airlines makes more on SABRE than it does on tickets**
- **Product Reviews, Advice = $$$**
  - (Likewise the DB of info adapters)
- **Optimal strategy for info access**
  - (Etzioni FOCS 96)

Meta-is-'betta'

- **Web Search**
- **Shopping**
- **Product Reviews**
- **Chat Finder**
- **Columnists (e.g. jokes, sports, ….)**
- **Email Lookup**
- **Event Finder**
- **People Finder**
- **Restaurant Reviews**
- **Job Listings**
- **Classifieds**
- **Apartment + Real Estate**

Personal Shopping Assistant

- **You Name the Product**
- **Agent Visits Stores, Review Sites...**
- **Makes Summary Report…**
- **If Requested, Buys Items for You**

Inevitable New Category

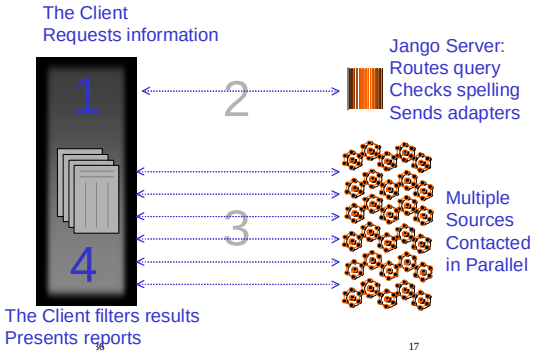
The Netbot Story: Part I

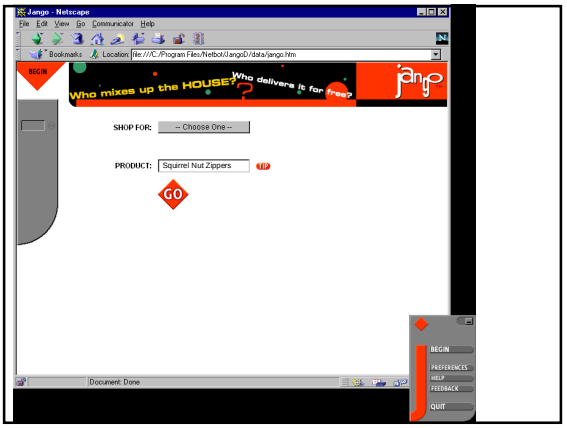
- **Bargain Finder (1995)**
  - Proof of concept; CD stores only; hardcoded
- **UW Shopbot Prototype (Jan '96)**
  - Focus on scalability to many stores, categories
- **Netbot Founded (May '96)**
  - Five Developers, 4 licenses, Seed funding

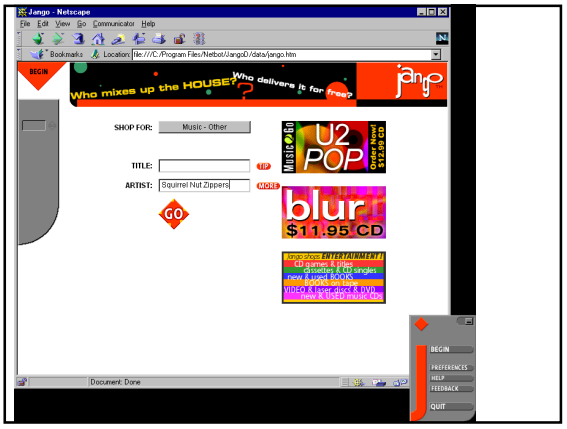
The Netbot Story: Continued

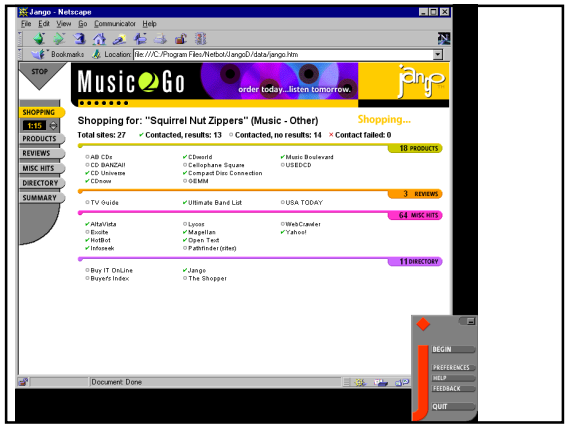
- **Hired Eric Zocher (Sept '96)**
  - Weld/Etzioni back to school
- **Second Round Financing (Jan '97)**
  - Netbot now at 25 people
- **Jango Client Launch**
  - Beta (April '97)
  - 1.0 (July '97)

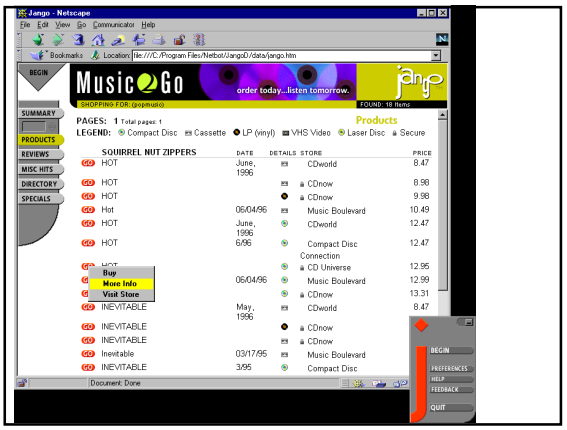
Jango Client Architecture



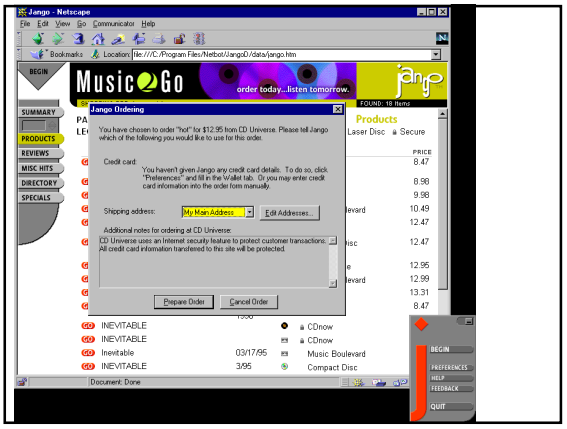


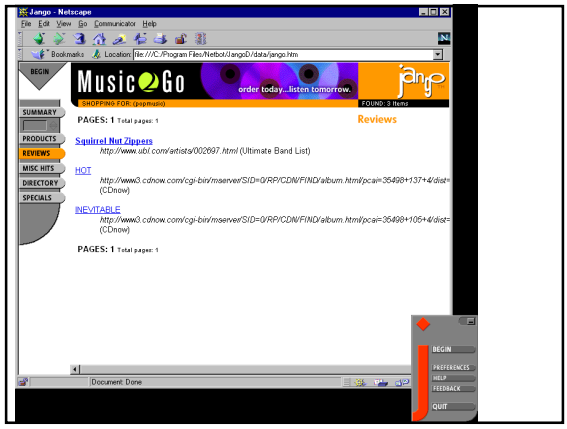

Jango - Netscape
Who mixes up the HOUSE? Who delivers it for free?
SHOP FOR: Music - Other
TITLE:
ARTIST: Squirrel Nut Zippers
GO
U2 POP
blur
$11.95 CD
Document: Done


Jango - Netscape
Music2Go
order today...listen tomorrow.
Shopping for: "Squirrel Nut Zippers" (Music - Other)
Shopping...
Total sites: 27 Contacted, results: 13 Contacted, no results: 14 Contact failed: 0
Document: Done


Jango - Netscape
Music2Go
order today...listen tomorrow.
PAGES: 1
LEGEND: Compact Disc Cassette LP (vinyl) VHS Video Laser Disc Secure
Products
SQUIRREL NUT ZIPPERS
Buy
More Info
Visit Store
Document: Done


Hot [ECD] * Album Info at CD Universe - Netscape
Squirrel Nut Zippers
Click Price to Put in Cart: $12.95 $9.15
CD Stock Status In Stock
Category ROCK/POP
Street Date Jun 4, 1996
Engineer Brian Paulson; Mike Napolitano
Studio/Live Studio
Label Mammoth Records
Catalog# 137
Album Notes: HOT contains both a full audio album and a CD-ROM program on one disc. The album is playable on any standard CD player; the computer program is playable on multimedia-equipped Macintosh and Windows systems. Squirrel Nut Zippers: Jim Mathus (vocals, guitar, tenor banjo, piano); Tom Maxwell (vocals, guitar, baritone saxophone, clarinet, resonator, chandelier); Katherine Whalen (vocals, banjo, barito
Mosher (guitar, alto & baritone saxophones, baritone ukulele, vocals); Don Raleigh (acoustic b
Phillips (drums, percussion, spoken vocals). Additional personnel: Andrew Bird (violin); Duke H
Document: Done


Jango - Netscape
Music2Go
order today...listen tomorrow.
Jango Ordering
You have chosen to order "hot" for $12.95 from CD Universe. Please tell Jango which of the following you would like to use for this order.
Credit card:
You haven't given Jango any credit card details. To do so, click "Preferences" and fill in the Wallet tab. Or you may enter credit card information into the order form manually.
Shipping address: My Main Address
Edit Addresses...
Additional notes for ordering at CD Universe:
CD Universe uses an Internet security feature to protect customer transactions. All credit card information transferred to this site will be protected.
Prepare Order
Cancel Order
Products
Document: Done


Jango - Netscape
Music2Go
order today...listen tomorrow.
PAGES: 1
Reviews
Squirrel Nut Zippers
http://www.ubl.com/artists/002697.html (Ultimate Band List)
HOT
(CDnow)
INEVITABLE
(CDnow)
PAGES: 1
Document: Done

Core Technology I:
Information Adapters

- **Low cost, flexible, "glue"**
  - Connects Jango to 100s of sites
- **Development & QA**
  - Proprietary language & compiler
  - Semiautomatic construction (learning)
  - Continuous monitoring

Core Technology II:
Query Routing

- **Selects WWW sites given query**
- **Limited natural language abilities**
- **Spell checking**
- **Taxonomic knowledge base**

Core Technology III:
Parallel Pull

- **Conventional Pull**
  - Typing a URL yields a single web page
- **Parallel Pull**
  - Typing high-level request yields info compiled from dozens of web pages
- **Parallel Aggregation Engine**
  - Efficient parallel access: aggregation, relevance, duplicate elimination

Consumer Benefits of Jango

- Comprehensive. **Knows about 100s of stores**
- Quick. **Contacts stores simultaneously**
- Current. **Gathers up-to-the-minute details**
- Convenient. **Creates custom shopping guide**
  - Product prices & availability...
  - Reviews...
  - Links to each online store...

Jango Benefits for Retailers

- **More Traffic.** Jango pulls from store with every query to a category.
- **Targeted Shoppers.** Jango users are looking for specific products.
- **One click buy.** Jango speeds sales process.
- **Reduced cost of customer acquisition.** Jango will catalyze more shopping
- **Shopping Reports.** What's selling

Jango Business Model

- **Give client away for free**
- **Get 1,000,000 users by Xmas '97**
- **Sell to merchants:**
  - Advertising (targeted users in act of buying!)
  - Real time promotions
  - Eventually aim for transaction cut

Reaction to Jango

- **Early adopters loved it**
- **Awards**
- **Substantial "buzz" in the press**
- **Two imitators (Sept '97)**
- **Contract with ATT Worldcom**
- **But only *30,000* downloads**

- **All efforts shifted to server implementation 8/97**

New Business Model

- **Shopping Infrastructure Company**
- **License Server to High-traffic Sites**
  - Search Engines: Yahoo, Excite, ...
  - ISPs: ATT Worldcom, AOL, MSN, …
  - Specialty sites: ESPN, ...
- **Split revenue from merchants**

Enter Excite

- **Very Aggressive Company**
  - #7 -> #2 search engine in 18 months
  - 26 Million pages views / day
- **Believed in Internet Shopping**
  - Already had leading shopping channel
  - More users/day than we got in 6 months
  - Sales force of 40
- **Acquires Netbot for $35 M**
  - Net Present Value of $300M in @Home stock

Copycats

- **Junglee**
  - Snagged Yahoo contract for Xmas '97
  - Bought by Amazon for $185M in Aug '98
    - Net present Value of $1.11B
  - ZShops?
  - Eliminate technology?
- **C2B**
  - Bought by Inktomi for $90M in Sept '98
    - Net Present Value of $630M
- **Infospace**
  - Activeshopper June '99
- **MySimon**
  - Bought by Cnet for $700M in Jan '00
- **Cadabra**
  - Bought by Goto for $250M in Feb '00

The Future

- **Price Tracking**
- **Auctions**
- **Category Wizards**
  - E.g. cars, long distance carriers
- **Cross Selling**
  - E.g. batteries, concert tickets
- **Gift Advisor**
- **Virtual Registry**
- **Cross Vendor Loyalty Program**