CYBERMetrics '99

Tracing agents and other automatic sampling procedures for the World Wide Web

Isidro F. Aguillo
CINDOC-CSIC
isidro@cindoc.csic.es

Seventh International Conference on Scientometrics and Informetrics. Colima, 1999

Sampling problems in the Web

- **No database with near complete coverage**
  - Altavista 250 Mpp < estimated 600 Mpp
  - links not validated (moved, no longer available, not accesible)
  - not pertinent answers (noise)
    - inadequate search capabilities
    - no metadata or invalid
  - irregular behaviour (saturation dependent): “sampling procedures”
- **Quality of the resources**
  - formal quality
  - value of the contents

First Generation possibilities

- **Search engines**
  - Advanced search facilities
    - No export facilities + limits
    - limited clustering capabilities (excl. Northern Light)
    - irregular behaviour in hibrid (boolean+field delimited) searching
- **Metasearchers**
  - Clustering
    - *Husky Search*

      http://huskysearch.cs.washington.edu/huskysearch
    - *Inference Find 2*

      http://infindv2.infind.com/infind.rdc

Towards Second Generation

- **Multisearchers**
  - *Copernic 99 3.01*
  - *Quest 99 2.03*
  - *Mata Hari 1.11*
  - *MetaSearch 2.2*
- **Advantages and disadvantages**
  - Low cost (shareware or free distribution)
  - Limited capabilities to export records
  - Small samples (200-300 records per engine)
  - Validation

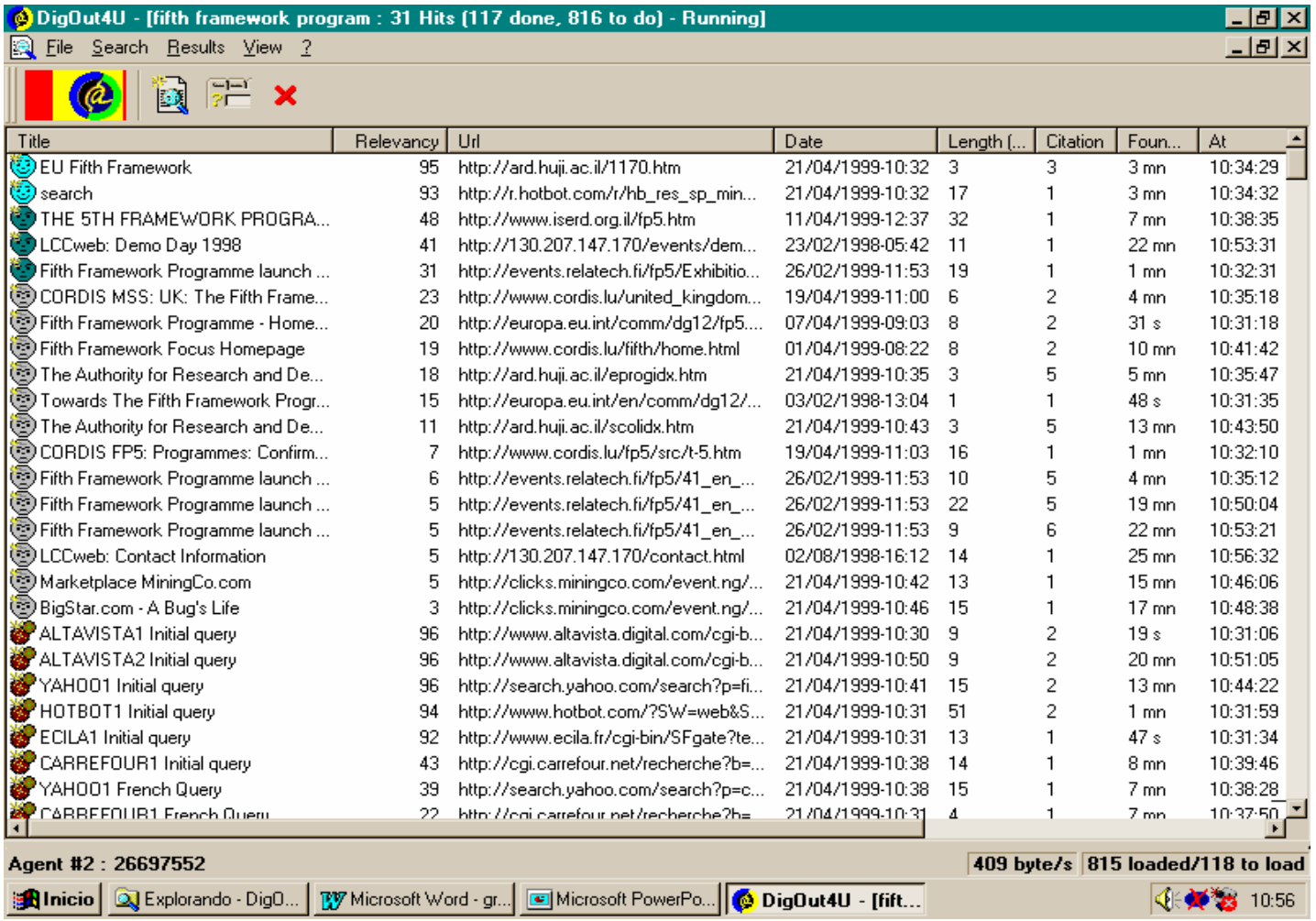
Another option

- **Tracing Agents**
  - DigOut4U 1.5
  - Cybot 2.42
  - Macrobot 3.03 Pro
  - WebBandit 3.60
- **Advantages and disadvantages**
  - Medium to high cost (some shareware)
  - Unlimited number of records to sample
  - Good capabilities to export records
    - **Access**
    - **HTML**

Tracing agents

- **Global ("virtual") capabilities**
  - Starting point
    - multisearcher
    - "seed/s", even some of the large indexes or search engine
  - Adding forbidden sites: "noise reduction"
  - Deep tracing level (or time limitation)
  - Search strategy
    - main terms ("root" recovery)
    - secondary terms (children and grandchildren sites)
    - differential weighting words and sites
  - Editing and exporting options
    - several formats

DigOut4U



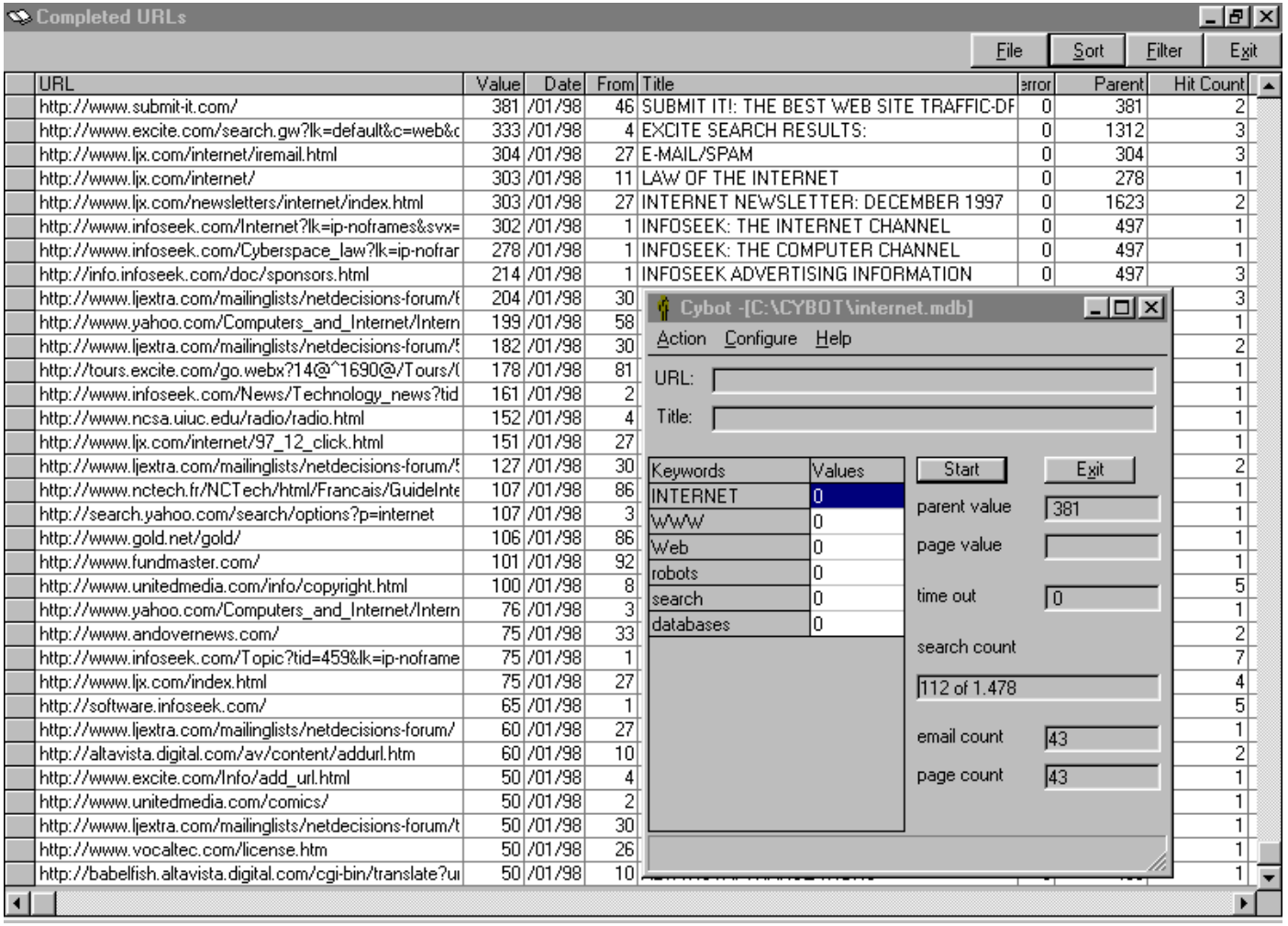
Title	Relevancy	Url	Date	Length (...	Citation	Foun...	At
EU Fifth Framework	95	http://ard.huji.ac.il/1170.htm	21/04/1999-10:32	3	3	3 mn	10:34:29
search	93	http://r.hotbot.com/r/hb_res_sp_min...	21/04/1999-10:32	17	1	3 mn	10:34:32
THE 5TH FRAMEWORK PROGRA...	48	http://www.iserd.org.il/fp5.htm	11/04/1999-12:37	32	1	7 mn	10:38:35
LCCweb: Demo Day 1998	41	http://130.207.147.170/events/dem...	23/02/1998-05:42	11	1	22 mn	10:53:31
Fifth Framework Programme launch ...	31	http://events.relatech.fi/fp5/Exhibitio...	26/02/1999-11:53	19	1	1 mn	10:32:31
CORDIS MSS: UK: The Fifth Frame...	23	http://www.cordis.lu/united_kingdom...	19/04/1999-11:00	6	2	4 mn	10:35:18
Fifth Framework Programme - Home...	20	http://europa.eu.int/comm/dg12/fp5....	07/04/1999-09:03	8	2	31 s	10:31:18
Fifth Framework Focus Homepage	19	http://www.cordis.lu/fifth/home.html	01/04/1999-08:22	8	2	10 mn	10:41:42
The Authority for Research and De...	18	http://ard.huji.ac.il/eprogidx.htm	21/04/1999-10:35	3	5	5 mn	10:35:47
Towards The Fifth Framework Progr...	15	http://europa.eu.int/en/comm/dg12/...	03/02/1998-13:04	1	1	48 s	10:31:35
The Authority for Research and De...	11	http://ard.huji.ac.il/scolidx.htm	21/04/1999-10:43	3	5	13 mn	10:43:50
CORDIS FP5: Programmes: Confirm...	7	http://www.cordis.lu/fp5/src/t-5.htm	19/04/1999-11:03	16	1	1 mn	10:32:10
Fifth Framework Programme launch ...	6	http://events.relatech.fi/fp5/41_en_...	26/02/1999-11:53	10	5	4 mn	10:35:12
Fifth Framework Programme launch ...	5	http://events.relatech.fi/fp5/41_en_...	26/02/1999-11:53	22	5	19 mn	10:50:04
Fifth Framework Programme launch ...	5	http://events.relatech.fi/fp5/41_en_...	26/02/1999-11:53	9	6	22 mn	10:53:21
LCCweb: Contact Information	5	http://130.207.147.170/contact.html	02/08/1998-16:12	14	1	25 mn	10:56:32
Marketplace MiningCo.com	5	http://clicks.miningco.com/event.ng/...	21/04/1999-10:42	13	1	15 mn	10:46:06
BigStar.com - A Bug's Life	3	http://clicks.miningco.com/event.ng/...	21/04/1999-10:46	15	1	17 mn	10:48:38
ALTAVISTA1 Initial query	96	http://www.altavista.digital.com/cgi-b...	21/04/1999-10:30	9	2	19 s	10:31:06
ALTAVISTA2 Initial query	96	http://www.altavista.digital.com/cgi-b...	21/04/1999-10:50	9	2	20 mn	10:51:05
YAHOO1 Initial query	96	http://search.yahoo.com/search?p=fi...	21/04/1999-10:41	15	2	13 mn	10:44:22
HOTBOT1 Initial query	94	http://www.hotbot.com/?SW=web&S...	21/04/1999-10:31	51	2	1 mn	10:31:59
ECILA1 Initial query	92	http://www.ecila.fr/cgi-bin/SFgate?te...	21/04/1999-10:31	13	1	47 s	10:31:34
CARREFOUR1 Initial query	43	http://cgi.carrefour.net/recherche?b=...	21/04/1999-10:38	14	1	8 mn	10:39:46
YAHOO1 French Query	39	http://search.yahoo.com/search?p=c...	21/04/1999-10:38	15	1	7 mn	10:38:28
CARREFOUR1 French Query	22	http://cgi.carrefour.net/recherche?b=...	21/04/1999-10:31	4	1	7 mn	10:37:50

Seventh International Conference on Scientometrics and Informetrics. Colima, 1999

DigOut4U

- **Description**
  - expensive “suite” (English/French customisation)
  - Multisearcher “seeds” generator
    - Simultaneous search agents (standard 5)
  - Downloading options
- **Some results**
  - Powerful program
    - Huge number of results
    - High quality (pertinent sites)
  - Large to very large files generated
  - filtering domains not working well
  - confusing interface

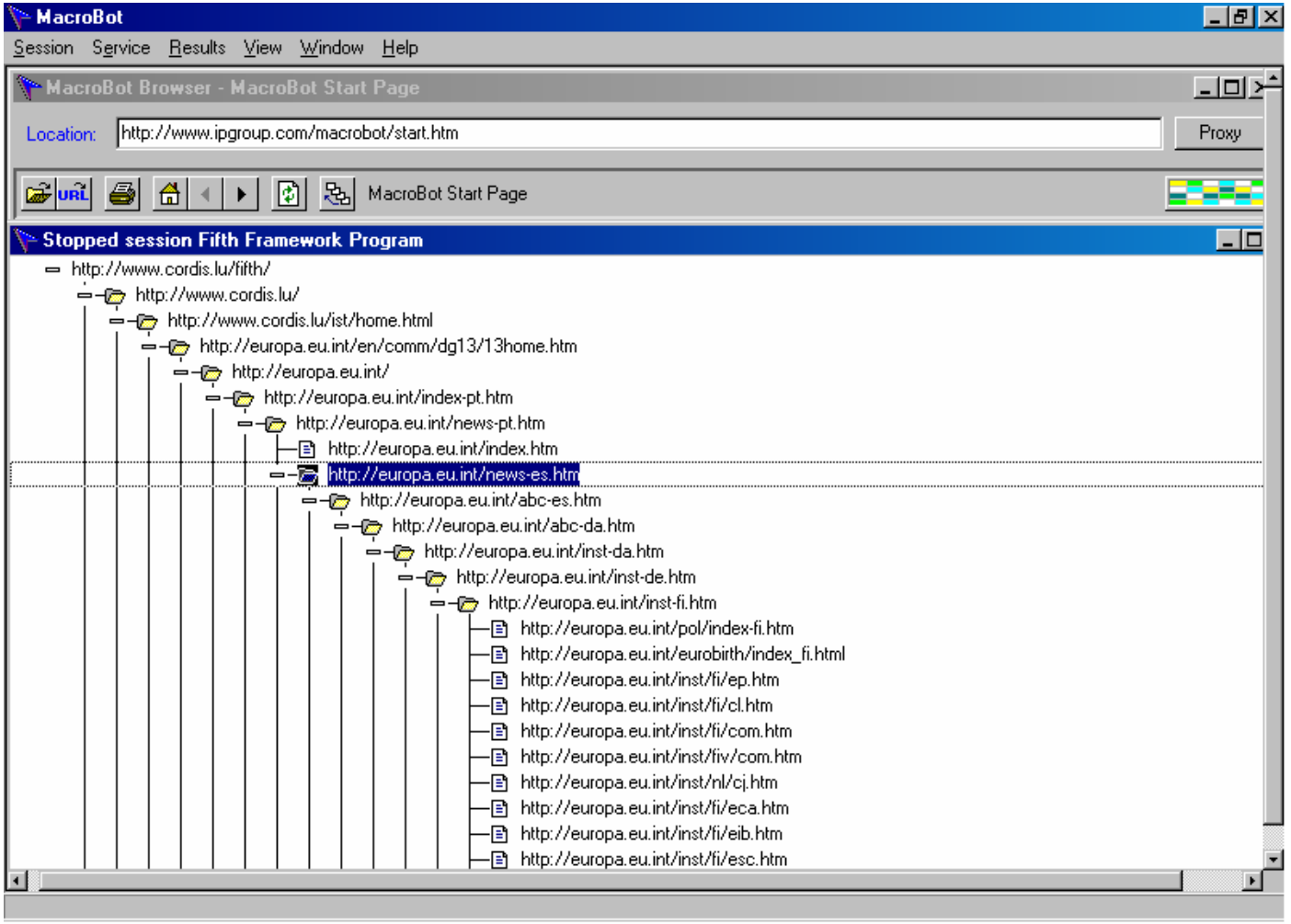
Cybot 2.42


Completed URLs
File
Sort
Filter
Exit
URL | Value | Date | From | Title | error | Parent | Hit Count
http://www.submit-it.com/ | 381 | /01/98 | 46 | SUBMIT IT!: THE BEST WEB SITE TRAFFIC-DF | 0 | 381 | 2
http://www.excite.com/search.gw?lk=default&c=web&c | 333 | /01/98 | 4 | EXCITE SEARCH RESULTS: | 0 | 1312 | 3
http://www.ljx.com/internet/iremail.html | 304 | /01/98 | 27 | E-MAIL/SPAM | 0 | 304 | 3
http://www.ljx.com/internet/ | 303 | /01/98 | 11 | LAW OF THE INTERNET | 0 | 278 | 1
http://www.ljx.com/newsletters/internet/index.html | 303 | /01/98 | 27 | INTERNET NEWSLETTER: DECEMBER 1997 | 0 | 1623 | 2
http://www.infoseek.com/Internet?lk=ip-noframes&svx= | 302 | /01/98 | 1 | INFOSEEK: THE INTERNET CHANNEL | 0 | 497 | 1
http://www.infoseek.com/Cyberspace_law?lk=ip-nofrar | 278 | /01/98 | 1 | INFOSEEK: THE COMPUTER CHANNEL | 0 | 497 | 1
http://info.infoseek.com/doc/sponsors.html | 214 | /01/98 | 1 | INFOSEEK ADVERTISING INFORMATION | 0 | 497 | 3
http://www.ljextra.com/mailinglists/netdecisions-forum/t | 204 | /01/98 | 30 | 3
http://www.yahoo.com/Computers_and_Internet/Intern | 199 | /01/98 | 58 | 1
http://www.ljextra.com/mailinglists/netdecisions-forum/t | 182 | /01/98 | 30 | 2
http://tours.excite.com/go.webx?14@^1690@/Tours/t | 178 | /01/98 | 81 | 1
http://www.infoseek.com/News/Technology_news?tid | 161 | /01/98 | 2 | 1
http://www.ncsa.uiuc.edu/radio/radio.html | 152 | /01/98 | 4 | 1
http://www.ljx.com/internet/97_12_click.html | 151 | /01/98 | 27 | 1
http://www.ljextra.com/mailinglists/netdecisions-forum/t | 127 | /01/98 | 30 | 2
http://www.nctech.fr/NCTech/html/Francais/GuideInte | 107 | /01/98 | 86 | 1
http://search.yahoo.com/search/options?p=internet | 107 | /01/98 | 3 | 1
http://www.gold.net/gold/ | 106 | /01/98 | 86 | 1
http://www.fundmaster.com/ | 101 | /01/98 | 92 | 1
http://www.unitedmedia.com/info/copyright.html | 100 | /01/98 | 8 | 5
http://www.yahoo.com/Computers_and_Internet/Intern | 76 | /01/98 | 3 | 1
http://www.andovernews.com/ | 75 | /01/98 | 33 | 2
http://www.infoseek.com/Topic?tid=459&lk=ip-noframe | 75 | /01/98 | 1 | 7
http://www.ljx.com/index.html | 75 | /01/98 | 27 | 4
http://software.infoseek.com/ | 65 | /01/98 | 1 | 5
http://www.ljextra.com/mailinglists/netdecisions-forum/ | 60 | /01/98 | 27 | 1
http://altavista.digital.com/av/content/addurl.htm | 60 | /01/98 | 10 | 2
http://www.excite.com/Info/add_url.html | 50 | /01/98 | 4 | 1
http://www.unitedmedia.com/comics/ | 50 | /01/98 | 2 | 1
http://www.ljextra.com/mailinglists/netdecisions-forum/t | 50 | /01/98 | 30 | 1
http://www.vocaltec.com/license.htm | 50 | /01/98 | 26 | 1
http://babelfish.altavista.digital.com/cgi-bin/translate?u | 50 | /01/98 | 10 | 1
Cybot -[C:\CYBOT\internet.mdb]
Action
Configure
Help
URL:
Title:
Keywords
Values
INTERNET
WWW
Web
robots
search
databases
Start
Exit
parent value
381
page value
time out
0
search count
112 of 1.478
email count
43
page count
43

Cybot 2.42

- **Description**
  - Complex system to configure
    - seeds
    - weights
  - MS Access as database engine
- **Some results**
  - Large number of analysed urls
    - low to very low number (ratio) of selected sites
    - pertinent to very pertinent
  - Confusing weighting system (more time is needed to master)
  - simple Interface

Macrobot 3.03

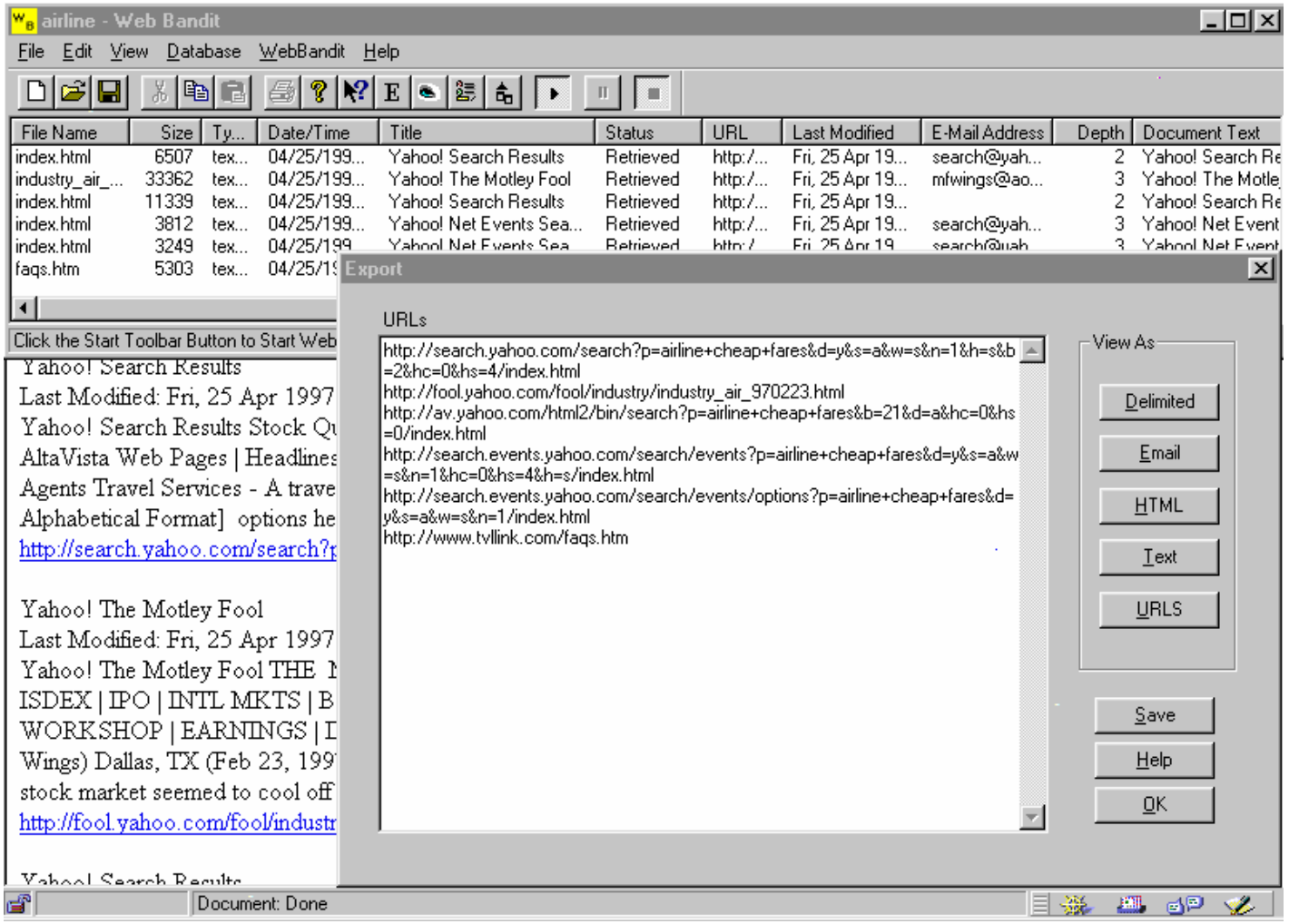


Seventh International Conference on Scientometrics and Informetrics. Colima, 1999

Macrobot 3.03

- **Description**
  - Programmable by “scripts”
    - –preconfigured
    - –macros
  - Only one agent (collapsible)
  - Exporting capabilities (Access)
- **Some results**
  - A beta version that generates problems with large number of records
  - Powerful, with filters devoted to extract data according several criteria (email addresses, by example)
  - Results are very pertinent, but they are mainly from derived from a “core” of sites

WebBandit 3.60



WebBandit 3.60

- Description
  - A bunch of possibilities
    - multisearch
    - download
    - indexing
    - extractor
  - Exporting options: html; MS Access
- Some results
  - The worst of the series according to the results: high level of noise and not-valid sites
  - The records are very rich with a good number of “fields”, but there are better search agents and indexers

Conclusions and recomendations

- Many of the search engines and recovery tools are not suitable to make samples of web resources for quantitative analysis
- The increasing size of the web and its hypertextual nature offer opportunities for a novel approach
- A new generation of recovering tools involving tracing hypertext links from selected sites are very promising
  - Offering capabilities to automate tasks
  - Extracting large samples of high pertinence
  - Ready to use in standard database formats
  - Selecting additional resources by indirect recovery via hypertext links