Automatic Gathering of Newspaper Articles on Internet Abuse from the Internet

Results of Phase 1 of the OSILIA Project

JRC, 20 December 2000

European Commission – Joint Research Centre
Institute for Systems, Informatics and Safety (ISIS)

Stefan Scheer (RMDS - Anti-fraud Information Management)
Ralf Steinberger (RMDS - Anti-fraud Information Management)
Paul Henshaw (RIT - Web Technologies)
Neil Mitchison (RIT - Dependable Software Applications)
Agenda

- Introduction to OSILIA (NM)
- Technical Goals of the OSILIA Project (RS)
- Commercial Tools and Services (RS)
- Functionality of the Software (SS)
- Storage and Retrieval of the Data (PH)
- Evaluation of the OSILIA-specific Achievements (NM)
- Future Work and Conclusions (RS)
“Cybercrime” is a Vague Term

- attacks on computers
- attacks on the Internet
- computer-based fraud (internet-based or not)
- crime happening to pass over Internet
- crimes which can be detected through Internet
- things people disapprove of (racism, spam, invasion of privacy, spying)

OSILIA uses the first four categories as a taxonomy for events
High-level concern about vulnerabilities (e.g. Tampere)

Commission services interested
JAI & INFSO, also MARKT, ENTER, SANCO…

Major policy questions, important technological aspects

JRC’s mission => JRC interest

IPTS and ISIS interested

Does Cybercrime happen? If so, how much and what?

- anecdotal evidence
- a few widely-reported incidents
- iceberg?

Are there lots of other incidents reported? Newspapers/Internet
Technical Goals

- Gather relevant open source information daily
classify and store it,
using a software agent and other programs

- Search for information with a mid- or long-term interest,
as opposed to ad hoc internet search to satisfy spontaneous needs

- Concentrate on newspaper articles (in the wider sense)
because they are manually pre-filtered information.
We want to have control over what we download
CyberAlert (www.cyberalert.com)
- scans English language sites and Usenet groups
- makes them available via a database on the internet
- or forwards the information found
- Approximate cost: setup fee of a few hundred USD and 360 USD per topic and month (~5,000 USD / year and topic)

BBC Monitoring (www.monitor.bbc.co.uk)
- gathers and translates published news from 100 languages into English

Reuters, ...
Commercial Crawlers

- IBM’s Intelligent Miner for Text
  www.software.ibm.com/iminer/fortext/: > 25 K€

- Verity’s Agent Server
  www.verity.com/products/agentserver/: 50 - 90 KUSD

- Excalibur’s Internet Spider
  www.excalibur.com: “similar price”

- Autonomy’s Knowledge Update
  www.autonomy.com: “six-digit sum in GBP” ~ minimum of 165 K€

- Tenmax’ Teleport Pro
  www.tenmax.com/teleport/pro/: 40 USD

- ...
Using Commercial Crawlers for OSILIA

- No such budget available in OSILIA and deadlines for delivery too tight
- Usually more functionality than what we need
- Often do not exactly satisfy our requirements
  - *Teleport Pro* does not allow Boolean queries
  - *Verity* and *Excalibur* only index, but do not store downloaded documents
- Advantage of own software development:
  - full control and adaptation to our needs
  - we are currently satisfied with the functionality of our crawler
- Usage of commercial software poses the same problems and still requires our own programs for filtering, cleaning, etc
General Approach

- **crawler**: obtain information from the web
- **convert**: convert document in text format
- **strip / cut**: extract the “kernel”
- **compare**: throw away multiple copies
- **throw away “index”** files
- **analyse**: look for search words
- **categorise**: define document relevance with respect to categories
- **assign keywords**: to facilitate document search
- **store**
- **query**
Hacker Hits on Pentagon Computers Up 10% This Year

By Walter Pincus
Washington Post Staff Writer
Saturday, December 9, 2000;
Page A08

The Defense Department suffered more than 22,000 electronic attacks on its computer systems in 1999 and about 14,000 in the first seven months of this year, the Pentagon's chief information officer said this
Hacker Hits on Pentagon Computers Up 10% This Year (washingtonpost.com)

For Help, press F1
Hacker Hits on Pentagon Computers Up 10% This Year (washingtonpost.com)

The Defense Department suffered more than 22,000 electronic attacks on its computer systems in

The vast majority of those attacks were either harmless or caused only petty harassment, but i

Pentagon officials said that, to the best of their knowledge, the Department of Defense's clas

The department was able to make an accurate count of the number of attacks for the first time

In 1999, the Pentagon detected 22,144 attempts to probe, scan, hack into, infect with viruses

So far this year, officials said, the number of attacks is up approximately 10 percent, and th

In an interview, Money predicted that the number of attacks is only "going to increase."

"A majority of the attacks [that cause damage] come through vulnerabilities in existing softw

Although the Pentagon is "putting more and more effort into testing" off-the-shelf software an

"On a lot of these [programs], we don't know where the code is written," he said.

Many of the vulnerabilities are unintentional, but some appear to be "trapdoors" deliberately

As a result, the official added, "we are not buying such off-the-shelf products in our most se

The Pentagon's cybersecurity problem is enormous. The Defense Department has roughly 10,000 co

In August, Congress put an additional $163 million for computer security into the fiscal 2001

The "seminal event" that awakened the Pentagon to its computer security problems occurred in E

Those attacks, which came as preparations were underway for a possible military operation agai

Military computer administrators had been warned about the weakness that the California hacker

© 2000 The Washington Post Company
DAILY OPINION

STRATEGICALLY SPEAKING

Political fallout
Small businesses should expect the 2000 election to affect technology choices and costs.

By Maggie Biggs

DAILY ARTICLES

ReleVents  Ask Molly  Computer Advisor

Are dot-com unions needed?  English to IT  Try downloading from mirrors

TODAY’S NEWS

FCC Airwave Auction Crests $1 Billion In Net Bids
A bevy of wireless telecommunications companies aiming to broaden their national 'footprints' drove net bidding in the FCC's

E-BUSINESS

Penguins running wild
E-VIEWS
Linux, the little OS that could, is poised to make a splash in e-business.

E-biz platforms open
E-BUSINESS FEATURE
up
There’s still a place for proprietary tools in e-commerce architecture, but Web-savvy companies are adopting open standards.

Unix vs. Linux
E-BUSINESS STEP-BY-STEP
Unix still makes sense for large-scale applications, but more and more e-businesses are powered by the OS’s younger sibling.

FEATURES

2001: an earthbound essay
A handful of ComputerUser's top products prove worth their weight.
General Methodology

Download

m web sites

html2text

analyse

filtered out

strip / cut

categorise

assign keywords

n databases

query & display

application specific parameters about web sites & files (sources)

application specific parameters about keywords, destinations & thresholds
Destination = apbnews
ApplyQuery = bodyOnly
FileTypes = text html
Roaming = directory
TrimTags = all
Traversal = breadthfirst
DownloadPages = yes
DownloadImages = no
SearchDepth = 2
Timeout = 20
Retries = 1
AgentLifetime = 70
MaxFiles = 400
MaxSize = 10
RetainOnlyLastVersion = no
StartingUrl = www.apbnews.com/newscenter/internetcrime/
Query = /internet|cyber|\bweb\b|\bWWW\b|computer|electronic|digital/ix &&
     /theft|criminal|trojan|pa?edophil|abuse|\bintru|virus|\bhack|fraud|spoof/ix
Destination = Paperball
Roaming = directory
Traversal = breadthfirst

Query = /internet\|cyber[^\s]*\|computer|ele\{ck\}troni|digital\|ix
StartingUrl =
  www.paperball.de/service/paperball.fcg?action=query&pg=detail&fmt=.&r=kriminal%2A+diebstahl+trojan%2A+missbrauch+p%E4do%2A&q=kriminal%2A+diebstahl+trojan%2A+missbrauch+p%E4do%2A&stq=71&d0=&d1=&what=german_web&rankedBy=date&categories=all&papers=all
Exclusions =
  www.paperball.de/service/paperball.fcg?action=query&categories=pol&papers=all&tp=f&pg=detail&r=&fmt=.&d0=&d1=&q=;
  www.paperball.de/service/paperball.fcg?action=query&categories=wir&papers=all&tp=f&pg=detail&r=&fmt=.&d0=&d1=&q=;
  www.paperball.de/service/paperball.fcg?action=query&categories=pol&papers=all&tp=f&pg=detail&r=&fmt=.&d0=&d1=&q=;
  www.paperball.de/service/paperball.fcg?action=query&categories=spo&papers=all&tp=f&pg=detail&r=&fmt=.&d0=&d1=&q=;
  www.paperball.de/service/paperball.fcg?action=query&categories=kun&papers=all&tp=f&pg=detail&r=&fmt=.&d0=&d1=&q=;
  www.paperball.de/service/paperball.fcg?action=query&categories=lok&papers=all&tp=f&pg=detail&r=&fmt=.&d0=&d1=&q=;
  www.paperball.de/service/paperball.fcg?action=query&categories=bun&papers=all&tp=f&pg=detail&r=&fmt=.&d0=&d1=&q=;
  www.tvtoday.de/paperball;
  www.paperball.de/service/voyeur-paperball-queries.fcg;
  www.paperball.de/service/guestbook-paperball.fcg;
  www.bol.de/cec/cstage?ecaction=boideeplink&amp;template=boiquicksearchresults.de.htm\startnum=1\increment=20\query_type=clue&referrer=011001960001\keyword1_entered=kriminal+diebstahl+trojan+missbrauch+p%E4ndo
Basic Search Words (Crawler)

- Boolean formula: \((a_1 \mid a_2 \mid \ldots \mid a_n) \AND (b_1 \mid b_2 \mid \ldots \mid b_m)\)

- Ab computer|b internet|b web|b WWW|b server|b cyber|b on [-]line|b e[-]mail|b net|b network|b password|b credit.card|b program|b firewall|b encryption/gi & &/ab us|b attack|b assault|b assail|b hack|war|b crack|b crim|b virus|b worm|b trojan.horse|b pae?doph|b terror|b intrusion|b child.pornogra/gi

- Ab computer|b internet|b web|b WWW|b server|b cyber|b on [-]line|b e[-]mail|b net|b netz|b password|b kredit[- ]?karte|b program|b firewall|verschlüsselung/gi & &/mi[sß]+brauch|angriff|attache|bhack|krimin|vir(us|en)|bwurm|btrojan|bpädoph|terror|beindri ng|b kind|porno/gi
English Language News Sites

Apbnews ([www.apbnews.com/newscenter/internetcrime](http://www.apbnews.com/newscenter/internetcrime))


Guardian ([www.guardian.co.uk](http://www.guardian.co.uk))

Newsbbc ([news.bbc.co.uk](http://news.bbc.co.uk))

Quicklinks ([www.qlinks.net/quicklinks/crypto.htm](http://www.qlinks.net/quicklinks/crypto.htm), [www.qlinks.net/quicklinks/comcrime.htm](http://www.qlinks.net/quicklinks/comcrime.htm), [www.qlinks.net/quicklinks/content.htm](http://www.qlinks.net/quicklinks/content.htm), [www.qlinks.net/quicklinks/rating.htm](http://www.qlinks.net/quicklinks/rating.htm)); meta news site

Der Standard (derstandard.at)

Kurier (www2.kurier.at)

Die Presse (194.158.136.155/textversion.taf): text version

Frankfurter Rundschau (www.f-r.de/fr)

Süddeutsche Zeitung
   (www.sueddeutsche.de/aktuell/?section=showAll&myTM=text): text version, no exclusions

Tagesspiegel (www.tagesspiegel.de): roaming allowed

PC Welt (www.pcwelt.de/index.html)

Spiegel (www.spiegel.de)

Paperball: newspaper articles search engine, various parameter files

Neue Zürcher Zeitung (newsticker.nzz.ch): few news, frequent changes; flat hierarchy, no exclusions

Tagesanzeiger (tagesanzeiger.ch/ta, tagesanzeiger.ch/computer)
OSILIA-Type Categories

● Attacks on computers:
  viruses, some Trojan Horses

● Computer-based fraud:
  “cracking”, credit card stripping,
  “man-in-the-middle” attacks,
  attacks on (financial) encryption

● Attacks on the internet itself:
  worms, some Trojan Horses, DoS attacks,
  hacking of web sites

● Crime using the internet:
  paedophilia, terrorism, bomb-making instructions,
  incitement
Special Search Words & Incrementing Counters

- FBI, Scotland Yard (each counter + 1)
- virus, worm, hoax (computer c. +2)
- melissa, love bug (computer c. +2)
- privacy, cookie, download (intern. +1)
- worm (internet counter +3)
- fraud [only once!] (fraud counter +2)
- crack, credit card, man in the middle (fraud counter +4)
- child, abus (crime counter +1)
- paedoph (crime counter +5)
Difficulties Encountered

- Dynamics of site names: good maintenance needed
- Dynamics / variety of web page contents: how to parameterize the stripping / cutting program? Maintenance of exclusions
- Multiply downloaded copies: hard to avoid
- ‘compare’ program too time-consuming
Possible Improvements

- Start with HTML code: structure!

- Meta news sites: how to fit into the general approach? How to handle “next” button?

- “Dynamic” search depth: in order to avoid “index” files
1. **Aided by Internet, Identity Theft Soars** *(Web Article)*

**Keywords:** cohn social number rend company san personal diego crime cyberspace enforcement issuer privacy consumer apartment information online fraud name docusearch security usa ligand chariton thief card internet theft credit identity bradbury

Aided by Internet, Identity Theft Soars By TIMOTHY L. O&acute;BRIEN aw enforcement authorities are becoming increasingly worried about a sudden, sharp rise in the incidence of identity theft, the outright pilfering of people's personal information for use in obtaining credit cards, loans a...

2. **Extortion Hack Raises Doubt of Online Security** *(Web Article)*

**Keywords:** info vendor wilkins posts company internet back-end computer information paar maxius com extortion hack e-commerce say database apbnews noack security fahhn web online
card credit

A mirrored image of the Web page NEW YORK (APBnews.com) -- A failed extortion plot that resulted in the online posting of a music store's customer credit card database has raised questions about cybersecurity just weeks after a record-setting holiday shopping season. Claiming to have hack...

3. **Was Bill Gates' Credit Card Number Online?** *(Web Article)*

**Keywords:** posts cd gate extortion incident database expert say sohn nbci xcom post e-mail web hack hacker maes security apbnews noack com number maxius posting online site credit card curador e-commerce

NEW YORK (APBnews.com) -- A self-proclaimed teenage "cracker" who claims to have gotten into the database of an e-commerce site has posted hundreds of credit card numbers online -- including one he says belongs to Microsoft Chairman Bill Gates.
Browse Articles (1)

OSILIA

browse categories

[ list 1416 Web Articles now ]

subject type

crime using internet, (130) attacks on Internet, (327)
computer based fraud, (459) attacks on computers, (500)

top 100 keywords

internet, (857) e-commerce, (89) information, (57)
com, (623) website, (88) technology, (57)
online, (522) attack, (87) suspect, (56)
web, (433) cnet, (85) cnn, (55)
hacker, (414) http, (79) virtuell, (55)
computer, (392) yahoo, (78) agency, (54)
virus, (291) iloveyou, (78) parent, (52)
site, (281) kid, (77) wap, (52)
ms, (247) entertainment, (77) zeitung, (52)
news, (226) cyber, (77) criminal, (51)
fi, (212) vermisches, (77) consumer, (50)
e-mail, (207) raquo, (77) copyright, (50)
bug, (182) porn, (76) pdt, (50)
love, (173) de, (76) program, (49)
email, (168) info, (75) investigation, (48)
software, (160) spiegel, (75) arrest, (48)
submissions, (157) computerworld, (74) story, (48)
Refine Search by Keyword

[ documents ] [ events ] [ folders ] [ contacts ] [ news ]

browse categories

list 459 Web Articles regarding computer based fraud now

top 100 keywords
internet, (349)  cnet, (48)  bbc, (31)
com, (281)    child, (46)    information, (29)
web, (190)    entertainment, (46)  uk, (29)
hacker, (167)  user, (44)    msnbc, (29)
online, (162)  crime, (43)    expert, (28)
computer, (158)  network, (43)  datum, (28)
site, (144)    block, (43)    privacy, (27)
fi, (104)      peacefire, (43)  technology, (27)
news, (95)    love, (42)    iloveyou, (27)
say, (81)      cybercrime, (39)  sex, (27)
email, (73)    noack, (39)    cyber, (26)
 virus, (76)    edit, (38)    firm, (26)
 html, (73)    score, (38)    electronic, (25)
software, (70)  censorware, (38)  zdnets, (25)
apnews, (70)    slashdot, (38)  enforcement, (24)
www, (67)      repo, (38)    yahoo, (24)
bug, (67)      copa, (38)    copyright, (24)
security, (66)  censorship, (38)  agency, (23)
porn, (65)      friday, (38)    story, (23)
 fraud, (63)    censor, (38)    reno, (23)
read, (63)      insightful, (28)  retnier, (23)
Aided by Internet, Identity Theft Soars

By TIMOTHY L. O’BRIEN

Law enforcement authorities are becoming increasingly worried about a sudden, sharp rise in the incidence of identity theft, the outright pilfering of people’s personal information for use in obtaining credit cards, loans and other goods. While identity theft is hardly new -- con men have been doing it for ages -- law enforcement officials and consumer advocates say the Internet is making identity theft one of the signature crimes of the digital era. Any visitor to cyberspace can find Web sites selling all sorts of personal information and, with that information in hand, thieves can acquire credit, make purchases and even secure residences in someone else’s name.

Can Privacy Be Protected Online? The Social Security Administration said it had received more than 30,000 complaints about the misuse of Social Security numbers last year, most of which had to do with identity theft. That was up from about 11,000 complaints in 1998 and 7,868 complaints in 1997. The Social Security Administration attributes the rise to the ease with which Social Security data can be collected on the Internet. "The Internet allows these criminals to work anonymously and from all over the world," said Gregory Regan, head of the Secret Service’s financial crimes division. "We don’t think that people should be afraid of using the Internet, but all of this enthusiasm should be tempered with a little bit of caution." Kim Bradbury knows all too well the ease with which identity thieves operate. On a Friday afternoon last September, as she was feeding her 1-year-old daughter in her Northern California home, the kitchen phone rang. Bradbury dismissed the caller, a representative of the credit card giant First USA, as just another telemarketing huckster. "No, you have to listen!" Bradbury remembers the caller insisting. "Someone has applied for a credit card in your name and it doesn’t match other addresses we have for you." In fact, someone had already used the credit card -- acquired over the Internet without Bradbury’s knowledge -- to...
A Moderate Success

- Technical - successful
  - Material located
  - Filters effective
- “Hit rates”:
  - Material relevant - 95%
  - Article relevant - 75%
  - Non-overlapping (retained) - 25%
  - Innovative, important - 5%
- Question “Are there other incidents?” answered
- Answer negative
- Consistent with cuttings
Possibilities for Further Work

- Improve the prototype (better cleaning, avoid duplicates, …)
- Add more sources and languages
- Apply to further domains of interest
- Assist users in specifying their interest profiles
- Further document analysis: Extraction of information such as
  - names of people and organisations
  - geographical references, etc.
- Add document similarity measure to database
- Cross-language keyword assignment and document comparison
- Visualisation of the document collection or subsets of it, using document maps
Conclusions

- OSILIA is one possible application domain. Other applications are planned (e.g. customs fraud-related subjects for OLAF)
- We are looking for further customers to finance this development

Applying the software to a new domain:
Setting up parameters for a specific interest profile will require several days or even weeks

- Usefulness of the software has to be decided case by case
- Precision is lower than human press clipping/classification, **but**
  - The range of sources is potentially very big
  - In the long term the automatic procedure is much cheaper
  - Results can be checked manually ➔ speeding up the manual work

- JRC Technical Note on the detailed methodology in early 2001