A Kosher Source of Ham

Nathan Friess
John Aycock

Department of Computer Science
University of Calgary
Canada

Building A Ham Corpus

- Hard To Publish Samples
  - Copyright Issues
  - Privacy Issues
- Some Goals
  - Realistic Samples
    - Ex: English words, grammar
  - Variety of Contexts
    - Technical writing, conversational

Related Work

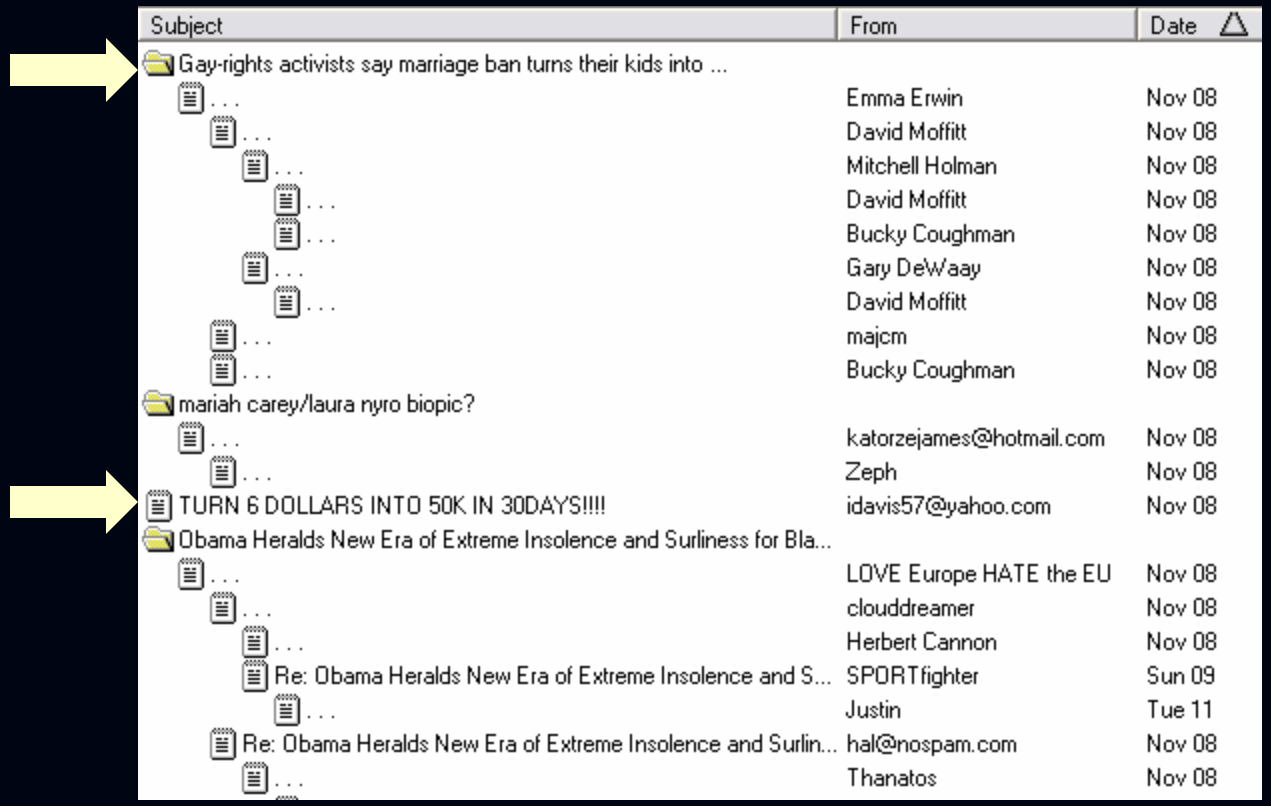
- Garcia, Hoepman, van Nieuwenhuizen (2004)
  - Simulation of legitimate email users, spammers, mailing lists
  - Gather ham text from Usenet
- But what about Usenet spam?

Usenet

- Public discussions
  - Contributing and viewing
- Organized into newsgroups
- “Typical” interaction:
  - One person posts
  - Many people reply to form a thread
- Underlying protocols:
  NNTP, RFC 1036 messages

From: Alice <alice@acme.com>
Newsgroups: rec.food.cooking
Subject: Re: Tasty chicken
Date: Sun, 13 Apr 2008 00:47:32 -0700 (PDT)
Message-ID: <abcd@acme.com>
References: <xyz123@googlegroups.com>

On Apr 12, 7:27 pm, Bob <bob@gmail.com> wrote:
>
>BBQ chicken is the best!
>
>Bob
>
I agree, especially on a hot summer day.
Alice



Subject	From	Date
Gay-rights activists say marriage ban turns their kids into ...		
. . .	Emma Erwin	Nov 08
. . .	David Moffitt	Nov 08
. . .	Mitchell Holman	Nov 08
. . .	David Moffitt	Nov 08
. . .	Bucky Coughman	Nov 08
. . .	Gary DeWaay	Nov 08
. . .	David Moffitt	Nov 08
. . .	majcm	Nov 08
. . .	Bucky Coughman	Nov 08
mariah carey/laura nyro biopic?		
. . .	katorzejames@hotmail.com	Nov 08
. . .	Zeph	Nov 08
TURN 6 DOLLARS INTO 50K IN 30DAYS!!!!	idavis57@yahoo.com	Nov 08
Obama Heralds New Era of Extreme Insolence and Surliness for Bla...		
. . .	LOVE Europe HATE the EU	Nov 08
. . .	clouddreamer	Nov 08
. . .	Herbert Cannon	Nov 08
Re: Obama Heralds New Era of Extreme Insolence and S...	SPORTfighter	Sun 09
. . .	Justin	Tue 11
Re: Obama Heralds New Era of Extreme Insolence and Surlin...	hal@nospam.com	Nov 08
. . .	Thanatos	Nov 08

Hypothesis

- *Do not* harvest all Usenet articles!
- Harvest REPLIES in threads
- A reply
  - Isn't just "Re: " in subject
  - Has a "References" header

From: Alice <alice@acme.com>
Newsgroups: rec.food.cooking
Subject: Re: Tasty chicken
Date: Sun, 13 Apr 2008 00:47:32 -0700 (PDT)
Message-ID: <abcd@acme.com>
References: <xyz123@googlegroups.com>

On Apr 12, 7:27 pm, Bob <bob@gmail.com> wrote:
>
>BBQ chicken is the best!
>
>Bob
>
I agree, especially on a hot summer day.
Alice

Experiments

- Gather articles from several newsgroups
- Train DSPAM on TREC 05 corpus
- Use DSPAM to classify replies, non-replies
- Manually verify some of DSPAM's results

Newsgroup Lists

- Custom list (38 groups)
  - Google “high traffic”
  - Hand-picked for variety of contexts
  - English
- NewsAdmin top-100 text list (77 groups)
  - Removed testing groups, job postings, spamtrap, net-abuse

Example Newsgroups

- alt.gossip.celebrities
- alt.religion.scientology
- comp.lang.python
- linux.kernel
- misc.invest.stocks
- rec.food.cooking
- rec.games.pinball
- rec.motorcycles
- sci.electronics.repair
- sci.math
- soc.retirement

Pre-processing

- Discard headers
  - Simulations only interested in bodies
- Replies: Remove quoted text
  - Quoted text will be classified separately

Results (Custom)

	Spam	Ham
Non-Replies	15,323 (16.6%)	76,996
Replies	5,299 (1.2%)	420,956

- Non-replies have 10x more spam

Manual Classification

- Definition of spam: conservative
- E.g.
  - Repeated postings, similar text / templates
  - Repeated words
  - Only a few random words and a URL
- Off-topic is not spam
- Selling goods is not spam
- If in doubt, not spam

Results (Manual, Custom)

Replies		DSPAM	
		Spam	Ham
Manual	Spam	33	5
	Ham	467	495

Non-Replies		DSPAM	
		Spam	Ham
Manual	Spam	419	155
	Ham	81	345

Results (Top 100)

	Spam	Ham
Non-Replies	27,282 (16.8%)	134,909
Replies	55,421 (5.8%)	895,618

- Non-replies have 3x more spam
  - Non-English groups problematic

A Reply: False Positive

Newsgroups: sci.math
Subject: Re: WHY HAS THIS SITE BECOME SUCH A SPAM TARGET???
Message-ID: <QbYQj.2$fO5.306@wagner.videotron.net>
Date: Sun, 27 Apr 2008 06:03:27 -0400

> (quoted text removed)

I don't know what ICP is. Google works hard on detecting "click fraud". If John and Jane sell floral arrangements through e-commerce and compete with each other, John can hire people to click on Jane's ads. That's assuming Jane has ads displayed on Web pages (search engines, blogs, etc.). One common arrangement is that Jane pays a few pennies when someone clicks on an ad for her products. The money is divided between the blogger (for example) and those who send "good" ads for the blogger (targeted to the blogger's readers).

…

A Non-Reply: False Negative

Newsgroups: alt.gossip.celebrities
Subject: SEXY SOFIE
Date: Wed, 11 Jun 2008 01:46:41 -0700 (PDT)
Message-ID: <00340094-bbca-4eab-96a4-
deeb2861cdec@w4g2000prd.googlegroups.com>

SEXY SOFIE
http://smilybaby.blogspot.com/2007/06/sexy-sofie.html
http://groups.yahoo.com/group/Enjoyment_Park/join

- And many more like it…

Limitations

- Spammers can use References header?
- Requires either:
  - Generate fake Message-IDs
    - Easy to correlate in Usenet client
  - Harvest real Message-IDs
    - Requires additional bandwidth
    - Spam will be buried in threads, not as visible

Conclusions

- Harvesting replies in Usenet is a good source of ham
  - If you can tolerate some noise
- Replies: 1 – 6 % spam
- Non-replies: 3x, 10x worse

A Kosher Source of Ham

Nathan Friess
John Aycock

Department of Computer Science
University of Calgary
Canada