Fighting Web Spam

C. Castillo, M. Sydow, J. Piskorski, D. Weiss

09/2007
Carlos Castillo
Yahoo! Research Barcelona, Spain

Marcin Sydow
Polish-Japanese Institute of Information Technology, Poland

Jakub Piskorski
Joint Research Centre of the European Commission, Italy

Dawid Weiss
Poznan University of Technology, Poland
Contents

1 Introduction

2 Reference Corpus & Features

3 New Experimental Results
Part I

Introduction

Marcin Sydow
1 Search Engines

2 Web Spam
The Web today

– the largest source of information
The Web today

- the largest source of information

size:
The Web today

- the largest source of information

size:

22.800.000.000 (WorldWideWebSize.com, 28 Aug 2007)
11.500.000.000 (A. Gulli, 2005)
The Web today

- the largest source of information

**size:**

22,800,000,000 (WorldWideWebSize.com, 28 Aug 2007)
11,500,000,000 (A. Gulli, 2005)

**content:**
The Web today

– the largest source of information

size:

22,800,000,000 (WorldWideWebSize.com, 28 Aug 2007)
11,500,000,000 (A. Gulli, 2005)

content:

over 100TB of text
+ multimedia
The Web today

– the largest source of information

size:

22,800,000,000 (WorldWideWebSize.com, 28 Aug 2007)
11,500,000,000 (A. Gulli, 2005)

content:

over 100TB of text
+ multimedia

Web population:
The Web today

– the largest source of information

size:

22,800,000,000 (WorldWideWebSize.com, 28 Aug 2007)
11,500,000,000 (A. Gulli, 2005)

content:

over 100TB of text
+ multimedia

Web population:

300,000,000 (Nielsen/NetRatings 2007)
700,000,000 unique users (comScore World Metrix, 2006.03)
Searching information – among the top Web activities

1(source: Alexa.com, August 2007)
Searching information – among the top Web activities

What are the 3 most popular Web sites today?¹

¹(source: Alexa.com, August 2007)
What are the 3 most popular Web sites today?\(^1\)

- Google.com
- Yahoo.com
- MSN.com

\(^1\)(source: Alexa.com, August 2007)
Searching information – among the top Web activities

What are the 3 most popular Web sites today?¹

- Google.com
- Yahoo.com
- MSN.com

search-focused portals

¹(source: Alexa.com, August 2007)
Why search engines?

– to make this ocean of information usable for humans
Why search engines?

– to make this ocean of information usable for humans

Search engines are the main gate to the Web, today
Why search engines?

- to make this ocean of information usable for humans

Search engines are the main gate to the Web, today

Facts:

256,000,000 people used a search engine in December 2006
(Nielsen/NetRatings, 2006)
Some available statistics

500.000.000 queries per day globally (after Google, 2005)
Some available statistics

500,000,000 queries per day globally (after Google, 2005)
Some available statistics

500,000,000 queries per day globally (after Google, 2005)

For a major global search engine it is:

- 250,000,000 queries daily,
- almost 3000 queries/sec over, 80TB textual corpus (say)
- each query must be served under 1 second...
Some available statistics

500,000,000 queries per day globally (after Google, 2005)

For a major global search engine it is:

- 250,000,000 queries daily,
- almost 3000 queries/sec over, 80TB textual corpus (say)
- each query must be served under 1 second...

“... the competitors are one click away...”
Search Engine Architecture

(after: “Searching the Web”, A. Arasu, et al.)
Search Engines – seemingly simple task

Return Web documents containing specified keywords
Search Engines – seemingly simple task

Return Web documents containing specified keywords

Modules:

- **Crawler**
  - follow links and collect documents

- **Repository**
  - store the docs – enable updates, access, persistence

- **Index**
  - record: which word in which document?

- **Ranking System**
  - which docs fit best to the users’ needs?
  - which docs are inherently valuable?

- **Presentation Module**
  - find a good form of result visualisation

- **Service**
  - process queries, find docs, present results
Crawler architecture

(after: “Mining the Web” S. Chakrabarti, Morgan-Kaufmann, 2003)
Ranking

An average query: **thousands of returned** documents

Average Human capability: **a few inspected** results
An average query: *thousands of returned* documents

Average Human capability: *a few inspected* results

How to select these *few out of thousands* for the beginning of the result list? – search engines’ *primary issue*
An average query: **thousands of returned** documents

Average Human capability: **a few inspected** results

How to select these **few out of thousands** for the beginning of the result list? – search engines’ *primary issue*

The **Ranking System** plays a **central role** in search quality
Search Engines  Web Spam

Ranking

An average query: **thousands of returned** documents

Average Human capability: **a few inspected** results

How to select these **few out of thousands** for the beginning of the result list? – search engines’ **primary issue**

The **Ranking System** plays a **central role** in search quality

Ranking systems existed in “classic” IR, before, but needed substantial adaptation to the needs of WWW. (search engine “revolution” AD 1998)
Ranking System

Influences the search quality (= mission-critical), kept secret

1. Assign a score to each document.
2. Sort docs in non-increasing order.

Factors used for computing the ranking:

- text analysis (doc’s content, URL, meta tags, etc.)
- anchor text analysis
- link analysis
- query log analysis
- traffic analysis
- user history analysis (personalisation)
Text-based Ranking – classic IR approach

A “bag of words” representation of text (document, query):

- A vector: keywords as dimensions, some statistics as coordinates

- TF-IDF (term freq. – inverted doc. freq.) or its variants

- Text-based ranking: vector similarity between query and document (dimensionality reduction (SVD, etc.), context-building, etc.)

Some drawbacks, but this model worked quite well for controlled textual document collections.
WWW-specific issues concerning text analysis

Classic IR techniques are faced with **Web-specific** issues:

- low quality mixed with high quality
- extreme diversity (versus homogeneity in classic IR)
- self-description problem
- noise, errors, etc.
- adversarial aspects – easy to spam
A Remedy – Link Analysis

Links represent a social aspect of Web publishing (to some extent).

A link from document $p$ to document $q$: a positive judgement

- the author of $p$ concerns $q$ as “valuable”,
  because it was chosen out of billions other documents to link to (except link nepotism).

A simplistic assumption, but works in mass.

Web users implicitly “assess” the Web documents.

Example: PageRank – a famous link-based ranking algorithm
Example: PageRank – Basic Idea of Authority Flow

1. each page has some **authority**
2. each page distributes its authority equally through links
3. the authority of a page is the authority flowing into this page
• simplified PageRank:

\[ R(p) = \sum_{i \in \text{IN}(p)} R(i) / \text{outDeg}(i), \]  

(1)
**PageRank Equations**

- **simplified PageRank:**

\[
R(p) = \sum_{i \in \text{IN}(p)} \frac{R(i)}{\text{outDeg}(i)},
\]

(1)

- introducing “dumping factor” \(d\) and “personalization vector” \(v(p)\):

\[
R(p) = (1 - d) \sum_{i \in \text{IN}(p)} \frac{R(i)}{\text{outDeg}(i)} + d \cdot v(p)
\]

(2)
PageRank Equations

- simplified PageRank:

\[
R(p) = \sum_{i \in \text{IN}(p)} \frac{R(i)}{\text{outDeg}(i)},
\]

(1)

- introducing “dumping factor” d and “personalization vector” \(v(p)\):

\[
R(p) = (1 - d) \sum_{i \in \text{IN}(p)} \frac{R(i)}{\text{outDeg}(i)} + d \cdot v(p)
\]

(2)

- simple “dangling-links” correction:

\[
R(p) = (1 - d) \sum_{i \in \text{IN}(p)} \frac{R(i)}{\text{outDeg}(i)} + d \cdot v(p) + (1 - d) v(p) \sum_{i \in \text{ZEROS}} R(i),
\]

(3)
PageRank – summary

PageRank, introduced in Google (1998), now patented in USA.

Most search engines apply similar algorithms, nowadays.

Properties:

1. A pioneer successful link-based ranking algorithm (also: HITS)
2. Quite immune to spamming
3. Gave birth to numerous variants:
   - personalized PageRank
   - Topic-sensitive PageRank (i.e. “dynamic” version)
   - Trust-Rank, and Anti-TrustRank, (SE spam combating)
   - extensions of the underlying random surfer model (e.g. RBS)
1 Search Engines

2 Web Spam
A bit of Web Economics…

What makes Search Engines survive?
A bit of Web Economics...  

What makes Search Engines survive?

search-based advertising – 97% of Web search revenues  


Main types:

- sponsored links (aside search results)
- contextual ads (placed on Web-sites)
Internet advertising revenues accounted for approximately 5.9 percent of total U.S. ad spending* in 2006, up from approximately 4.7 percent in 2005.

**U.S. Advertising Market-Media Comparisons—2006 ($ Billions)**

- **Direct Mail**: $55.7
- **Newspapers**: $51.2
- **TV Networks: Broadcast & Cable**: $39.9
- **TV Distribution**: $32.5
- **Magazines**: $24.6
- **Radio**: $20.8
- **Internet**: $16.9
- **Outdoor**: $6.8

*The total U.S. advertising market is estimated at approximately $285 billion, and includes other segments not charted here.

Sources: IAB Internet Ad Revenue Report; PricewaterhouseCoopers Global Entertainment and Media Outlook
Internet Advertising (USA, 2006)

Search-based ads take the major share (40%) – $6.76B
Web pages are accessed through search engines

1. Search engine **ranking** → Web page **visibility**
2. Web page **visibility** → **traffic** on the page
3. **traffic** on the page → **incomes**

Thus it is **incentive** today to **rank highly** in search engines!
What is Spam?

**Definition**
Web Spam (Search Engine Spam) is any manipulation of Web documents in order to mislead Search Engines to obtain **undeservedly high ranking**, without improving the “real” document information quality (for humans)

or (the extreme version):

**Definition**
Web Spam (Search Engine Spam) is anything that Web authors do only because Search Engines exist.

Web Spam is motivated economically: $16.9B \times 40\% = \$6.76B$ (in 2006)
Spam is destructive

Spam affects every-day life of Web community
Spam is destructive

Spam affects every-day life of Web community

- undermines mission and business of search engines
Spam is destructive

Spam affects every-day life of Web community

- undermines mission and business of search engines
- seriously deteriorates information search quality in the Web
Spam is destructive

Spam affects every-day life of Web community

- undermines mission and business of search engines
- seriously deteriorates information search quality in the Web

Combating Web spam is a primary issue not only for search engines.
Spam vs SEO

Not all actions taken in order to improve Web visibility of pages are regarded as spam.

- “white hat” techniques for improving Web page visibility exist (SEO)
- SE publish their guidelines in their “Terms of Service”
- There is a gray area in between, however...
Spam taxonomy

Two groups of techniques:
Spam taxonomy

Two groups of techniques:
- hiding techniques
Spam taxonomy

Two groups of techniques:

- hiding techniques
- boosting techniques
Spam taxonomy

Two groups of techniques:
- hiding techniques
- boosting techniques
Spam taxonomy

Two groups of techniques:

- hiding techniques
- boosting techniques

With regard to factors used in ranking algorithms:
Spam taxonomy

Two groups of techniques:
- hiding techniques
- boosting techniques

With regard to factors used in ranking algorithms:
- content-based techniques
Spam taxonomy

Two groups of techniques:
- hiding techniques
- boosting techniques

With regard to factors used in ranking algorithms:
- content-based techniques
- link-based techniques
Spam taxonomy

Two groups of techniques:
- hiding techniques
- boosting techniques

With regard to factors used in ranking algorithms:
- content-based techniques
- link-based techniques
- other
Spam techniques

• content-based
  • hidden text (size, color)
  • repetition
  • keyword stuffing/dilution
  • language-model-based (phrase stealing, dumping)
Spam techniques

- content-based
  - hidden text (size, color)
  - repetition
  - keyword stuffing/dilution
  - language-model-based (phrase stealing, dumping)

- link-based
  - “honey pot”
  - anchor-text spam
  - blog/wiki spam
  - link exchange
  - link farms
  - expired domains
Spam techniques

- content-based
  - hidden text (size, color)
  - repetition
  - keyword stuffing/dilution
  - language-model-based (phrase stealing, dumping)

- link-based
  - “honey pot”
  - anchor-text spam
  - blog/wiki spam
  - link exchange
  - link farms
  - expired domains

- other
  - cloaking
  - redirection
cheap car hire call center [details here] or complete our simple cheap car hire enquiry form [here] and we will call you back.

[Cheep Auto Rental] [Cheep Airport Parking] [Cheep Travel Insurance] [Cheep Foreign Currency]
[Cheep Flight Tickets] [Cheep Hotel Booking] [Cheep Rent A Car] [Cheep Package Holidays] [Cheep Rent A Car Service]

Indexed by Linksmania
Terms & Conditions. Privacy Policy.
cheepcar.co.uk copyright cheeptravel Limited© cheeptravel Limited© part of the DHD Group Limited

DISCOUNTED CAR HIRE IN THE UK. For the best deal on CHEAP car hire rental in the United Kingdom, visit our UK DISCOUNT SELF DRIVE feature. Guaranteed discount off normal self drive rates.
DISCOUNTED CAR HIRE IN THE UK. For the best deal on CHEAP car hire rental in the United Kingdom, visit our UK DISCOUNT SELF DRIVE feature. Guaranteed discount off normal self drive rates.
DISCOUNTED CAR HIRE IN THE UK. For the best deal on CHEAP car hire rental in the United Kingdom, visit our UK DISCOUNT SELF DRIVE feature. Guaranteed discount off normal self drive rates.
Texas hold 'em (or simply hold 'em or holdem) is the most popular of the community card poker games. It is the most popular poker variant played in casinos in the western United States, and its no-limit form is used in the main event of the World Series of Poker (abbreviated WSOP), widely recognized as the world championship of the game.

Seven-card stud is a poker variant. Until the recent increase in popularity of Texas hold 'em, Seven-card stud was the most popular poker variant in home games across the United States, and in casinos in the eastern part of the country.

Omaha hold 'em (or Omaha holdem or simply Omaha) is a community card poker game based on Texas hold 'em. It was originally created as a high-hand only game, but a high-low split variant called "Omaha eight-or-better" has also become popular.

Five-card draw is often the first poker variant learned by most players, and is very common in home games although it is now rare in casino and tournament play. The lowball variations make more interesting games and are more commonly played in casinos. Two to eight players can play.

Home Security Webpage

Ads by Google

Alarm Systems
Looking to find alarm systems? Visit our alarm systems guide.
OnlyAlarmSystems.com

Security Systems
Selected Security System Deals Find Exactly What You Want Today
www.Security-Systems.in

Centurion Wireless System
Panic Alarm System for Public Facilities and Courthouses.
www.stopenehtd.com

Uncategorized 22 Nov 2005 02:03 pm

Home security system - Separate Blasts Kill Nearly 100 in Iraq

Separate Blasts Kill Nearly 100 in Iraq
Washington Post - By Ellen Knickmeyer and Naseer Nouri
Washington Post Foreign Service Saturday, November 19, 2005; Page A01 BAGHDAD, Nov. AP) Video
Security Video Shows Huge ExplosionVideo from a security camera at the Hamra Hotel in Baghdad look at the fallen troops' home towns, ages, service categories and other

Rooftop girl's game of strip

Archived Entry

Post Date :
Tuesday, Nov 22nd, 2005 at 2:03 pm

Category :
Uncategorized

Do More :
You can trackback from your own site.

Ads by Google

Prevent Home Burglary
Home burglary is rampant. Read all about security systems.
www.for-the-touchdown

Security Industry News
Latest on CCTV, loss prevention, access control & more for pros
SecurityInfoWatch.com
Search engine?

Results 1-16 containing "sports book:"

1. **Place Your Bet with #1 Sports Betting Site Online**
   Kentucky Derby, NBA, MLB, NHL and all other sports betting and odds. Place a full range of bets on betting sportsbook in North America
   [http://www.sportsinteraction.com](http://www.sportsinteraction.com)

2. **AnteUp GamblingLinks.com - Safe Online Casinos**
   Links to safe and secure online casino gambling and sports betting including reviews, news and more!
   [http://gamblinglinks.com](http://gamblinglinks.com)

3. **Free Casino Bonuses. Links To The Best Casinos**
   Get $20 - $500 in Free Chips. Most popular casino games with great graphics. Play for fun or real money. Ll rules and strategy. Links to the Best Casinos
   [http://www.fastfreecash.net](http://www.fastfreecash.net)

4. **AnteUp GamblingLinks.com - Safe Online Casinos**
Fake search engine

Top Searches:
- Canadian Pharmacy
- Debt Consolidation
- Online Loan
- Diet
- Credit Reports
- Online Poker
- Xenical
- Buy Ionamin
- Diet Pills
- Online Craps
- DirecTV
- Life Insurance
- Dedicated Server
- Car Insurance
- Buy Phentermine
- Debt
- Weight Loss Pills
- Pay Day Loans
- Home Loan
- Refinance

Top Web Results

Results 1-16 containing "1293kasd132ka0sd1kj239asd123"

1. A Real Work At Home Business Opportunity!
   Free Home Business Match Up Service! We have helped 1000's of people make $5,000.
   http://gozing.directtrack.com/z/1198/CD2127/

2. Exotic Holiday - Find Your Love
   Exotic holiday is great way how to find love when you travel. Meet new people. Meet
   http://www.exotic-holiday.co.uk/

3. Image, Photo, Digital, Video and Movie Software
   Find quality image management & digital asset software for your business. Also see
   http://www.enterprise-software.co.uk/

4. Renting a Birthday Party Limousine is Sexy
   What better way to surprise your loved one on their special day than with a birthday
   http://partybusrental.info
"Normal" content in link farms

Website design, management, marketing and promotion

If you are searching for any of the following topics:

- Website design, management, marketing and promotion
- Website design, management, marketing and promotion resources
- Website design, management, marketing and promotion related topics
- Website design, management, marketing and promotion services

Look No further. You'll find it at Website design, management, marketing and promotion.

Website design, management, marketing and promotion is the key to your needs. You're one step ahead with Dry Media.

Website design, management, marketing and promotion brought to you by Dry Media, the leaders in this field.

At the Website design, management, marketing and promotion web site, you'll discover an easy to use, information packed source of data on Website design, management, marketing and promotion.

Click Here to Learn More about Website design, management, marketing and promotion.
**Cloaking**

**Search Engine**

**Search engine results page**

**User**

**Click on the result**

**Normal document**

**Buy viagra now!**

**Cloaking:**

different contents at the same URL
Redirects using Javascript

Simple redirect

```html
<script>
document.location="http://www.topsearch10.com/";
</script>
```

“Hidden” redirect

```html
<script>
var1=24; var2=var1;
if(var1==var2) {
    document.location="http://www.topsearch10.com/";
}
</script>
```
Problem: obfuscated code

Obfuscated redirect

```html
<script>
var a1="win", a2="dow", a3="loca", a4="tion.",
    a5="replace", a6=("http://www.top10search.com/");
var i,str="",
for(i=1;i<=6;i++)
{
    str += eval("a"+i);
}
eval(str);
</script>
```
Problem: really obfuscated code

Encoded javascript

```javascript
var s = "%5CBE0D%5C%05GDHJ_BDE%16...%04%0E";
var e = ", i;
eval(unescape(’s%eDunescape%28s%29%3Bfor...%3B’));
```

More examples: [Chellapilla and Maykov, 2007]
Fighting Spam

On the search engines’ side:

- Education (what is regarded spam and what is not)
- Spam detection
  - text-based (contents, URLs, meta-tags)
  - link-based (Trust-Rank, Anti-TrustRank, etc.)
  - language-model based (Language Model Disagreement method, etc.)
- Maintaining up-to-date black lists
- Recently ML-based
- Maintaining spam-reporting interfaces
- Punishment (excluding from index)

For researchers:

Very interesting applications of Data Mining/Information Retrieval.
Fighting Spam

On the search engines’ side:

- Education (what is regarded spam and what is not)
Fighting Spam

On the search engines’ side:

- Education (what is regarded spam and what is not)
- Spam detection
Fighting Spam

On the search engines’ side:

- Education (what is regarded spam and what is not)
- Spam detection
  - text-based (contents, URLs, meta-tags)
- link-based (Trust-Rank, Anti-TrustRank, etc.)
- language-model based (Language Model Disagreement method, etc.)
- maintaining up-to-date black lists
- recently ML-based
- Maintaining spam-reporting interfaces
- Punishment (excluding from index)

For researchers:

Very interesting applications of Data Mining/Information Retrieval.
Fighting Spam

On the search engines’ side:

- Education (what is regarded spam and what is not)
- Spam detection
  - text-based (contents, URLs, meta-tags)
  - link-based (Trust-Rank, Anti-TrustRank, etc.)
Fighting Spam

On the search engines’ side:

- Education (what is regarded spam and what is not)
- Spam detection
  - text-based (contents, URLs, meta-tags)
  - link-based (Trust-Rank, Anti-TrustRank, etc.)
  - language-model based (Language Model Disagreement method, etc.)
Fighting Spam

On the search engines’ side:

- Education (what is regarded spam and what is not)
- Spam detection
  - text-based (contents, URLs, meta-tags)
  - link-based (Trust-Rank, Anti-TrustRank, etc.)
  - language-model based (Language Model Disagreement method, etc.)
- maintaining up-to-date “black lists”
Fighting Spam

On the search engines’ side:

- **Education** (what is regarded spam and what is not)
- **Spam detection**
  - text-based (contents, URLs, meta-tags)
  - link-based (Trust-Rank, Anti-TrustRank, etc.)
  - language-model based (Language Model Disagreement method, etc.)
  - maintaining up-to-date “black lists”
- recently – ML-based
Fighting Spam

On the search engines’ side:

- Education (what is regarded spam and what is not)
- Spam detection
  - text-based (contents, URLs, meta-tags)
  - link-based (Trust-Rank, Anti-TrustRank, etc.)
  - language-model based (Language Model Disagreement method, etc.)
  - maintaining up-to-date “black lists”
  - recently – ML-based
- Maintaining spam-reporting interfaces
Fighting Spam

On the search engines’ side:

- Education (what is regarded spam and what is not)
- Spam detection
  - text-based (contents, URLs, meta-tags)
  - link-based (Trust-Rank, Anti-TrustRank, etc.)
  - language-model based (Language Model Disagreement method, etc.)
  - maintaining up-to-date “black lists”
  - recently – ML-based
- Maintaining spam-reporting interfaces
- Punishment (excluding from index)
Fighting Spam

On the search engines’ side:

• Education (what is regarded spam and what is not)
• Spam detection
  • text-based (contents, URLs, meta-tags)
  • link-based (Trust-Rank, Anti-TrustRank, etc.)
  • language-model based (Language Model Disagreement method, etc.)
  • maintaining up-to-date “black lists”
  • recently – ML-based
• Maintaining spam-reporting interfaces
• Punishment (excluding from index)
Fighting Spam

On the search engines’ side:

- Education (what is regarded spam and what is not)
- Spam detection
  - text-based (contents, URLs, meta-tags)
  - link-based (Trust-Rank, Anti-TrustRank, etc.)
  - language-model based (Language Model Disagreement method, etc.)
  - maintaining up-to-date “black lists”
  - recently – ML-based
- Maintaining spam-reporting interfaces
- Punishment (excluding from index)

For researchers:
Very interesting applications of Data Mining/Information Retrieval.
ML can help greatly

The struggle gets harder:

- There are **a lot of factors** used to compute search engine ranking

- There is an “**arms race**”:
ML can help greatly

The struggle gets harder:

- There are a *lot of factors* used to compute search engine ranking
- There is an “arms race”:
  - 1. spammers apply new deceptive technique
ML can help greatly

The struggle gets harder:

- There are **a lot of factors** used to compute search engine ranking

- There is an “**arms race**”:
  - spammers apply new deceptive technique
  - search engine improves the ranking system
ML can help greatly

The struggle gets harder:

- There are a lot of factors used to compute search engine ranking
- There is an “arms race”:
  1. spammers apply new deceptive technique
  2. search engine improves the ranking system
  3. spammers apply new deceptive technique
ML can help greatly

The struggle gets harder:

- There are a lot of factors used to compute search engine ranking

- There is an “arms race”:
  1. spammers apply new deceptive technique
  2. search engine improves the ranking system
  3. spammers apply new deceptive technique
  4. search engine improves the ranking system...
ML can help greatly

The struggle gets harder:

- There are a lot of factors used to compute search engine ranking

- There is an “arms race”:
  1. spammers apply new deceptive technique
  2. search engine improves the ranking system
  3. spammers apply new deceptive technique
  4. search engine improves the ranking system...
ML can help greatly

The struggle gets harder:

- There are **a lot of factors** used to compute search engine ranking

- There is an "**arms race**":
  1. spammers apply new deceptive technique
  2. search engine improves the ranking system
  3. spammers apply new deceptive technique
  4. search engine improves the ranking system...

**Machine Learning** approach recently applied to support Search Engines in combating Web spam
Part II

Reference Corpus & State of the Art

Carlos Castillo
Tools for dealing with Web Spam

Publicly Known

- Too hard to implement
- Does not work in practice
- Has not been tried yet

Publicly known and used in practice

Used by Search Engines

- Too detailed to explain
- Creates competitive advantage
- Does not work if revealed
Motivation

Fetterly [Fetterly et al., 2004] hypothesized that studying the distribution of statistics about pages could be a good way of detecting spam pages:

“in a number of these distributions, outlier values are associated with web spam”
Challenges: Machine Learning

Machine Learning Challenges:

- Instances are not really independent (graph)
- Learning with few examples
- Scalability
Challenges: Information Retrieval

Information Retrieval Challenges:

- Feature extraction: which features?
- Feature aggregation: page/host/domain
- Feature propagation (graph)
- Recall/precision tradeoffs
- Scalability
3 A Reference Collection

4 Link-based features

5 Content-based features

6 Using Links and Contents

7 SIGIR’07: Exploiting Topology
Data is really important

- It is dangerous for a search engine to provide labelled data for this
- Even if they do, it would never reflect a consensus
Assembling Process

- Crawling of base data
- Elaboration of the guidelines and classification interface
- Labeling
- Post-processing
Crawling of base data

**U.K. collection**

77.9 M pages downloaded from the .UK domain in May 2006 (LAW, University of Milan)
### U.K. collection

77.9 M pages downloaded from the .UK domain in May 2006 (LAW, University of Milan)

- Large seed of about 150,000 .uk hosts
- 11,400 hosts
- 8 levels depth, with \( \leq 50,000 \) pages per host
Classification interface
Labeling process

- We asked 20+ volunteers to classify entire hosts
Labeling process

- We asked 20+ volunteers to classify entire hosts
- Asked to classify normal / borderline / spam
Labeled process

- We asked 20+ volunteers to classify entire hosts
- Asked to classify normal / borderline / spam
- Do they agree? Mostly...
Agreement
Results

• Labels:

<table>
<thead>
<tr>
<th>Label</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>4,046</td>
<td>61.75%</td>
</tr>
<tr>
<td>borderline</td>
<td>709</td>
<td>10.82%</td>
</tr>
<tr>
<td>spam</td>
<td>1,447</td>
<td>22.08%</td>
</tr>
<tr>
<td>can not classify</td>
<td>350</td>
<td>5.34%</td>
</tr>
</tbody>
</table>

• Agreement:

<table>
<thead>
<tr>
<th>Category</th>
<th>Kappa</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.62</td>
<td>Substantial agreement</td>
</tr>
<tr>
<td>spam</td>
<td>0.63</td>
<td>Substantial agreement</td>
</tr>
<tr>
<td>borderline</td>
<td>0.11</td>
<td>Slight agreement</td>
</tr>
<tr>
<td>global</td>
<td>0.56</td>
<td>Moderate agreement</td>
</tr>
</tbody>
</table>
Result: first public Web Spam collection

- Public spam collection
Result: first public Web Spam collection

- Public spam collection
  - Labels for 6,552 hosts
Result: first public Web Spam collection

- Public spam collection
  - Labels for 6,552 hosts
  - 2,725 hosts classified by at least 2 humans
Result: first public Web Spam collection

- Public spam collection
  - Labels for 6,552 hosts
  - 2,725 hosts classified by at least 2 humans
  - 3,106 automatically considered normal (.ac.uk, .sch.uk, .gov.uk, .mod.uk, .nhs.uk or .police.uk)
Result: first public Web Spam collection

- Public spam collection
  - Labels for 6,552 hosts
  - 2,725 hosts classified by at least 2 humans
  - 3,106 automatically considered normal (.ac.uk, .sch.uk, .gov.uk, .mod.uk, .nhs.uk or .police.uk)
- [http://www.yr-bcn.es/webspam/](http://www.yr-bcn.es/webspam/)
Result: first public Web Spam collection

- Public spam collection
  - Labels for 6,552 hosts
  - 2,725 hosts classified by at least 2 humans
  - 3,106 automatically considered normal (.ac.uk, .sch.uk, .gov.uk, .mod.uk, .nhs.uk or .police.uk)
  - [http://www.yr-bcn.es/webspam/](http://www.yr-bcn.es/webspam/)
- Web Spam challenge
Result: first public Web Spam collection

- Public spam collection
  - Labels for 6,552 hosts
  - 2,725 hosts classified by at least 2 humans
  - 3,106 automatically considered normal (\texttt{.ac.uk}, \texttt{.sch.uk}, \texttt{.gov.uk}, \texttt{.mod.uk}, \texttt{.nhs.uk} or \texttt{.police.uk})
  - \url{http://www.yr-bcn.es/webspam/}

- Web Spam challenge
  - Track I: Information retrieval + Machine learning
Result: first public Web Spam collection

- Public spam collection
  - Labels for 6,552 hosts
  - 2,725 hosts classified by at least 2 humans
  - 3,106 automatically considered normal (.ac.uk, .sch.uk, .gov.uk, .mod.uk, .nhs.uk or .police.uk)
  - [http://www.yr-bcn.es/webspam/](http://www.yr-bcn.es/webspam/)
- Web Spam challenge
  - Track I: Information retrieval + Machine learning
  - Track II: Machine learning
Result: first public Web Spam collection

- Public spam collection
  - Labels for 6,552 hosts
  - 2,725 hosts classified by at least 2 humans
  - 3,106 automatically considered normal (.ac.uk, .sch.uk, .gov.uk, .mod.uk, .nhs.uk or .police.uk)
  - [http://www.yr-bcn.es/webspam/](http://www.yr-bcn.es/webspam/)

- Web Spam challenge
  - Track I: Information retrieval + Machine learning
  - Track II: Machine learning
Result: first public Web Spam collection

- Public spam collection
  - Labels for 6,552 hosts
  - 2,725 hosts classified by at least 2 humans
  - 3,106 automatically considered normal (.ac.uk, .sch.uk, .gov.uk, .mod.uk, .nhs.uk or .police.uk)
  - [http://www.yr-bcn.es/webspam/](http://www.yr-bcn.es/webspam/)

- Web Spam challenge
  - Track I: Information retrieval + Machine learning
  - Track II: Machine learning

- AIRWeb 2007 Workshop
Result: first public Web Spam collection

- Public spam collection
  - Labels for 6,552 hosts
  - 2,725 hosts classified by at least 2 humans
  - 3,106 automatically considered normal (.ac.uk, .sch.uk, .gov.uk, .mod.uk, .nhs.uk or .police.uk)
  - [http://www.yr-bcn.es/webspam/](http://www.yr-bcn.es/webspam/)

- Web Spam challenge
  - Track I: Information retrieval + Machine learning
  - Track II: Machine learning

- AIRWeb 2007 Workshop
- GraphLab 2007 Workshop
3 A Reference Collection

4 Link-based features

5 Content-based features

6 Using Links and Contents

7 SIGIR'07: Exploiting Topology
Link farms

Single-level link farms can be detected by searching groups of pages sharing their out-links [Gibson et al., 2005].
Single-level farms can be detected by searching groups of nodes sharing their out-links [Gibson et al., 2005]
Semi-streaming model

Handling large graphs:

- Memory size enough to hold some data per-node
- Disk size enough to hold some data per-edge
- A small number of sequential passes over the data
Link-Based Features

- Degree-related measures
- PageRank
- TrustRank [Gyöngyi et al., 2004]
- Truncated PageRank [Becchetti et al., 2006]
- Estimation of supporters [Becchetti et al., 2006]
Degree-Based

![Bar charts and line plots comparing Normal and Spam categories across different metrics such as Degree/Degree ratio of home page and Number of in-links.](image)
TrustRank Idea

World Wide Web

“Trusted” Nodes

Suspicious
Estimated relative non-spam mass $d = 0.59$
Areas below the curves are equal if we are in the same strongly-connected component.
Areas below the curves are equal if we are in the same strongly-connected component.
Neighbors: spam
Neighbors: normal
Bottleneck number

\[ b_d(x) = \min_{j \leq d} \left\{ \frac{|N_j(x)|}{|N_{j-1}(x)|} \right\} \]. Minimum rate of growth of the neighbors of \( x \) up to a certain distance. We expect that spam pages form clusters that are somehow isolated from the rest of the Web graph and they have smaller bottleneck numbers than non-spam pages.
Probabilistic counting

Count bits set to estimate supporters

Propagation of bits using the "OR" operation

Target page

Becchetti et al., 2006 shows an improvement of ANF algorithm [Palmer et al., 2002] based on probabilistic counting [Flajolet and Martin, 1985].
[Becchetti et al., 2006] shows an improvement of ANF algorithm [Palmer et al., 2002] based on probabilistic counting [Flajolet and Martin, 1985]
3 A Reference Collection

4 Link-based features

5 Content-based features

6 Using Links and Contents

7 SIGIR’07: Exploiting Topology
Content-Based Features

Most of the features reported in [Ntoulas et al., 2006]

- Number of word in the page and title
- Average word length
- Fraction of anchor text
- Fraction of visible text
- Compression rate
- Corpus precision and corpus recall
- Query precision and query recall
- Independent trigram likelihood
- Entropy of trigrams

More about this in the last part of the talk
Content-based features (entropy related)

\( T = \{(w_1, p_1), \ldots, (w_k, p_k)\} \) the set of trigrams in a page, where trigram \( w_i \) has frequency \( p_i \).

Features:

- Entropy of trigrams \( H = - \sum_{w_i \in T} p_i \log p_i \)
- Also, compression rate, as measured by bzip
Content-based features (related to popular keywords)

$F$ set of most frequent terms in the collection
$Q$ set of most frequent terms in a query log
$P$ set of terms in a page

Features:

- Corpus “precision” \[ \frac{|P \cap F|}{|P|} \]
- Corpus “recall” \[ \frac{|P \cap F|}{|F|} \]
- Query “precision” \[ \frac{|P \cap Q|}{|P|} \]
- Query “recall” \[ \frac{|P \cap Q|}{|Q|} \]
Average word length

Figure: Histogram of the average word length in non-spam vs. spam pages for $k = 500$. 
Corpus precision

Figure: Histogram of the corpus precision in non-spam vs. spam pages.
Query precision

Figure: Histogram of the query precision in non-spam vs. spam pages for $k = 500$. 
3 A Reference Collection

4 Link-based features

5 Content-based features

6 Using Links and Contents

7 SIGIR’07: Exploiting Topology
## Cost-sensitive decision tree with bagging

Bagging of 10 decision trees, asymmetrical costs.

<table>
<thead>
<tr>
<th>Cost ratio</th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rate</td>
<td>65.8%</td>
<td>66.7%</td>
<td>71.1%</td>
<td>78.7%</td>
<td>84.1%</td>
</tr>
<tr>
<td>False positive rate</td>
<td>2.8%</td>
<td>3.4%</td>
<td>4.5%</td>
<td>5.7%</td>
<td>8.6%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.712</td>
<td>0.703</td>
<td>0.704</td>
<td><strong>0.723</strong></td>
<td>0.692</td>
</tr>
</tbody>
</table>
## Link- and content-based features

<table>
<thead>
<tr>
<th></th>
<th>Both</th>
<th>Link-only</th>
<th>Content-only</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rate</td>
<td>78.7%</td>
<td>79.4%</td>
<td>64.9%</td>
</tr>
<tr>
<td>False positive rate</td>
<td>5.7%</td>
<td>9.0%</td>
<td>3.7%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.723</td>
<td>0.659</td>
<td>0.683</td>
</tr>
</tbody>
</table>
3 A Reference Collection

4 Link-based features

5 Content-based features

6 Using Links and Contents

7 SIGIR’07: Exploiting Topology
General hypothesis

Pages topologically close to each other are more likely to have the same label (spam/nonsspam) than random pairs of pages.
General hypothesis

Pages topologically close to each other are more likely to have the same label (spam/nonspam) than random pairs of pages.

Pages linked together are more likely to be on the same topic than random pairs of pages [Davison, 2000]
Topological dependencies: in-links

Histogram of fraction of spam hosts in the in-links

- $0 = \text{no in-link comes from spam hosts}$
- $1 = \text{all of the in-links come from spam hosts}$
Topological dependencies: out-links

Histogram of fraction of spam hosts in the out-links

- 0 = none of the out-links points to spam hosts
- 1 = all of the out-links point to spam hosts
Idea 1: Clustering

Classify, then cluster hosts, then assign the same label to all hosts in the same cluster by majority voting
Idea 1: Clustering (cont.)

Initial prediction:
Idea 1: Clustering (cont.)

Clustering:
Idea 1: Clustering (cont.)

Final prediction:
## Idea 1: Clustering – Results

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Without bagging</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positive rate</td>
<td>75.6%</td>
<td>74.5%</td>
</tr>
<tr>
<td>False positive rate</td>
<td>8.5%</td>
<td>6.8%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.646</td>
<td>0.673</td>
</tr>
<tr>
<td><strong>With bagging</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positive rate</td>
<td>78.7%</td>
<td>76.9%</td>
</tr>
<tr>
<td>False positive rate</td>
<td>5.7%</td>
<td>5.0%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.723</td>
<td>0.728</td>
</tr>
</tbody>
</table>

☑ Reduces error rate
Idea 2: Propagate the label

Classify, then interpret “spamicity” as a probability, then do a random walk with restart from those nodes.
Idea 2: Propagate the label (cont.)

Initial prediction:
Idea 2: Propagate the label (cont.)

Propagation:
Idea 2: Propagate the label (cont.)

Final prediction, applying a threshold:
Idea 2: Propagate the label – Results

<table>
<thead>
<tr>
<th>Classifier without bagging</th>
<th>Baseline</th>
<th>Fwds.</th>
<th>Backwds.</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rate</td>
<td>75.6%</td>
<td>70.9%</td>
<td>69.4%</td>
<td>71.4%</td>
</tr>
<tr>
<td>False positive rate</td>
<td>8.5%</td>
<td>6.1%</td>
<td>5.8%</td>
<td>5.8%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.646</td>
<td>0.665</td>
<td>0.664</td>
<td>0.676</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier with bagging</th>
<th>Baseline</th>
<th>Fwds.</th>
<th>Backwds.</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rate</td>
<td>78.7%</td>
<td>76.5%</td>
<td>75.0%</td>
<td>75.2%</td>
</tr>
<tr>
<td>False positive rate</td>
<td>5.7%</td>
<td>5.4%</td>
<td>4.3%</td>
<td>4.7%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.723</td>
<td>0.716</td>
<td>0.733</td>
<td>0.724</td>
</tr>
</tbody>
</table>
Idea 3: Stacked graphical learning

- Meta-learning scheme [Cohen and Kou, 2006]
- Derive initial predictions
- Generate an additional attribute for each object by combining predictions on neighbors in the graph
- Append additional attribute in the data and retrain
Idea 3: Stacked graphical learning (cont.)

- Let $p(x) \in [0..1]$ be the prediction of a classification algorithm for a host $x$ using $k$ features
- Let $N(x)$ be the set of pages related to $x$ (in some way)
- Compute
  \[ f(x) = \frac{\sum_{g \in N(x)} p(g)}{|N(x)|} \]
- Add $f(x)$ as an extra feature for instance $x$ and learn a new model with $k + 1$ features
Initial prediction:

\[ y_a = p(x_{a1} \ldots x_{ak}) \]

\[ y_b = p(x_{b1} \ldots x_{bk}) \]

\[ y_c = p(x_{c1} \ldots x_{ck}) \]
Idea 3: Stacked graphical learning (cont.)

Computation of new feature:

\[ y_a = p(x_{a1} \ldots x_{ak}) \]
\[ z_a = f(y_b) \]
\[ y_c = p(x_{c1} \ldots x_{ck}) \]
\[ z_c = f(y_b) \]

\[ y_b = p(x_{b1} \ldots x_{bk}) \]
\[ z_b = f(y_a, y_c) \]
Idea 3: Stacked graphical learning (cont.)

New prediction with \( k + 1 \) features:

\[
\begin{align*}
    y_a &= p(x_{a1} \ldots x_{ak}) \\
    z_a &= f(y_b) \\
    y'_a &= p'(x_{a1} \ldots x_{ak}, z_a) \\
    y_c &= p(x_{c1} \ldots x_{ck}) \\
    z_c &= f(y_b) \\
    y'_c &= p'(x_{c1} \ldots x_{ck}, z_c) \\
    y_b &= p(x_{b1} \ldots x_{bk}) \\
    z_b &= f(y_a, y_c) \\
    y'_b &= p'(x_{b1} \ldots x_{bk}, z_b)
\end{align*}
\]
### Idea 3: Stacked graphical learning - Results

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Avg. of in</th>
<th>Avg. of out</th>
<th>Avg. of both</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rate</td>
<td>78.7%</td>
<td>84.4%</td>
<td>78.3%</td>
<td>85.2%</td>
</tr>
<tr>
<td>False positive rate</td>
<td>5.7%</td>
<td>6.7%</td>
<td>4.8%</td>
<td>6.1%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.723</td>
<td>0.733</td>
<td>0.742</td>
<td><strong>0.750</strong></td>
</tr>
</tbody>
</table>

✔ Increases detection rate
Idea 3: Stacked graphical learning x2

And repeat ...

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>First pass</th>
<th>Second pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rate</td>
<td>78.7%</td>
<td>85.2%</td>
<td>88.4%</td>
</tr>
<tr>
<td>False positive rate</td>
<td>5.7%</td>
<td>6.1%</td>
<td>6.3%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.723</td>
<td>0.750</td>
<td><strong>0.763</strong></td>
</tr>
</tbody>
</table>

✓ Significant improvement over the baseline
Part III

New Experimental Results

Jakub Piskorski, Marcin Sydow, Dawid Weiss
New results

- Linguistic features
- IDEA 1: simple addition of linguistic features
- IDEA 2: pruning incomplete data
- IDEA 3: selecting good “pure” hosts
- Summary
Why linguistic features?

- Using linguistic and language features such as language diversity, complexity, expressivity, immediacy, uncertainty and emotional consistency turned to have discriminatory potential for deception detection [Zhou et al., 2004].
- In previous research linguistic features not extensively exploited for web spam detection.
- Explore prevalence of spam relative to linguistic features in WEB-SPAM-2006UK corpus.
How to measure?

- Complexity: *average number of: sentences, clauses, noun phrases.*
- Diversity: *lexical diversity, content word diversity.*
- Expressivity: *preference of specific part-of-speech categories to others.*
- Non-immediacy: *self-reference, passive voice, generalizing terms.*
Linguistic features

- **Length** = total number of tokens (word-like units)
- **Lexical diversity** = \( \frac{\text{number of different tokens}}{\text{total number of tokens}} \)
- **Lexical validity** = \( \frac{\text{number of tokens which constitute valid word forms}}{\text{total number of potential word forms}} \)
- **Text-like fraction** = \( \frac{\text{total number of potential word forms}}{\text{total number of tokens}} \)
- **Emotiveness** = \( \frac{\text{number of adjectives and adverbs}}{\text{number of nouns and verbs}} \)
- **Self-referencing** = \( \frac{\text{number of 1st-person pronouns}}{\text{total number of pronouns}} \)
- **Passive voice** = \( \frac{\text{number of verb phrases in passive voice}}{\text{total number of verb phrases}} \)
Computing linguistic features

- Only for the “summary” of the WEB-SPAM-2006UK corpus (< 400 pages per host), 64GB.

- Utilized Corleone (Core Linguistic Entity Extraction), developed at JRC, and LingPipe (www.alias-i.com/lingpipe).

- 14.36% of pages had no “textual” content.
IDEA 1

Just add the linguistic features to the attribute set.
Idea 1: Just linguistic features

<table>
<thead>
<tr>
<th></th>
<th>linguistic features</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>with</td>
<td>without</td>
</tr>
<tr>
<td>instances</td>
<td>8,411</td>
<td>8,411</td>
</tr>
<tr>
<td>attributes</td>
<td>287</td>
<td>280</td>
</tr>
<tr>
<td>classified correctly</td>
<td>7,666 91.14%</td>
<td>7,687 91.39%</td>
</tr>
<tr>
<td>missclassified</td>
<td>745 8.85%</td>
<td>724 8.60%</td>
</tr>
</tbody>
</table>

• The results are not much different.
## Idea 1: Just linguistic features

With linguistic features:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP</th>
<th>FP</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.970</td>
<td>0.435</td>
<td>0.946</td>
<td>0.970</td>
<td>0.958</td>
</tr>
<tr>
<td>undecided</td>
<td>0.091</td>
<td>0.010</td>
<td>0.162</td>
<td>0.091</td>
<td>0.116</td>
</tr>
<tr>
<td>spam</td>
<td>0.525</td>
<td>0.033</td>
<td>0.615</td>
<td>0.525</td>
<td>0.566</td>
</tr>
</tbody>
</table>

Without linguistic features:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP</th>
<th>FP</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.970</td>
<td>0.415</td>
<td>0.949</td>
<td>0.970</td>
<td>0.959</td>
</tr>
<tr>
<td>undecided</td>
<td>0.108</td>
<td>0.010</td>
<td>0.186</td>
<td>0.108</td>
<td>0.137</td>
</tr>
<tr>
<td>spam</td>
<td>0.552</td>
<td>0.033</td>
<td>0.629</td>
<td>0.552</td>
<td>0.588</td>
</tr>
</tbody>
</table>

Figures in red are “better”.

### New results

Features

IDEA 1

IDEA 2

IDEA 3

Summary
Prune the input by removing records with missing values. Rerun the experiments with and without linguistic attributes.
Idea 2: prune records with missing values

<table>
<thead>
<tr>
<th></th>
<th>linguistic features</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>with</td>
<td>without</td>
<td></td>
</tr>
<tr>
<td>instances</td>
<td>6,644</td>
<td>6,644</td>
<td></td>
</tr>
<tr>
<td>attributes</td>
<td>287</td>
<td>280</td>
<td></td>
</tr>
<tr>
<td>classified correctly</td>
<td>6,016</td>
<td>90.54%</td>
<td>6,009</td>
</tr>
<tr>
<td>missclassified</td>
<td>628</td>
<td>9.45%</td>
<td>635</td>
</tr>
</tbody>
</table>

• Not much improvement (difference so small it is most likely statistically insignificant).
Idea 2: prune records with missing values

With linguistic features:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP</th>
<th>FP</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.958</td>
<td>0.343</td>
<td>0.954</td>
<td>0.958</td>
<td>0.956</td>
</tr>
<tr>
<td>undecided</td>
<td>0.112</td>
<td>0.019</td>
<td>0.119</td>
<td>0.112</td>
<td>0.115</td>
</tr>
<tr>
<td>spam</td>
<td>0.608</td>
<td>0.039</td>
<td>0.622</td>
<td>0.608</td>
<td>0.615</td>
</tr>
</tbody>
</table>

Without linguistic features:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP</th>
<th>FP</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.958</td>
<td>0.348</td>
<td>0.954</td>
<td>0.958</td>
<td>0.956</td>
</tr>
<tr>
<td>undecided</td>
<td>0.105</td>
<td>0.019</td>
<td>0.113</td>
<td>0.105</td>
<td>0.109</td>
</tr>
<tr>
<td>spam</td>
<td>0.601</td>
<td>0.039</td>
<td>0.616</td>
<td>0.601</td>
<td>0.608</td>
</tr>
</tbody>
</table>
IDEA 3

Choose only “pure” hosts (for which class decision was univocal).
Rerun the experiments with and without linguistic attributes.
The notion of a “spam host” is quite vague, inter-judge classification agreement is not perfect.

Selecting representative spam/ not spam records by filtering univocally-classified examples;

- 1049 NNN hosts,
- 391 SS hosts,
- 57 BB hosts,
- (no SSS or BBB examples in the original data).

The above gives a total of 1497 pure hosts used as input.
Idea 3: “pure” hosts

<table>
<thead>
<tr>
<th></th>
<th>linguistic features</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>with</td>
<td>without</td>
<td></td>
</tr>
<tr>
<td>instances</td>
<td>1,497</td>
<td>1,497</td>
<td></td>
</tr>
<tr>
<td>attributes</td>
<td>287</td>
<td>280</td>
<td></td>
</tr>
<tr>
<td>classified correctly</td>
<td>1,328</td>
<td>1,330</td>
<td></td>
</tr>
<tr>
<td></td>
<td>88.71%</td>
<td>88.84%</td>
<td></td>
</tr>
<tr>
<td>missclassified</td>
<td>169</td>
<td>167</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11.28%</td>
<td>11.15%</td>
<td></td>
</tr>
</tbody>
</table>
Idea 3: “pure” hosts

With linguistic features:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP</th>
<th>FP</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.949</td>
<td>0.107</td>
<td>0.954</td>
<td>0.949</td>
<td>0.952</td>
</tr>
<tr>
<td>undecided</td>
<td>0.193</td>
<td>0.042</td>
<td>0.155</td>
<td>0.193</td>
<td>0.172</td>
</tr>
<tr>
<td>spam</td>
<td>0.821</td>
<td>0.055</td>
<td>0.840</td>
<td>0.821</td>
<td>0.831</td>
</tr>
</tbody>
</table>

Without linguistic features:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP</th>
<th>FP</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.950</td>
<td>0.103</td>
<td>0.956</td>
<td>0.950</td>
<td>0.953</td>
</tr>
<tr>
<td>undecided</td>
<td>0.175</td>
<td>0.041</td>
<td>0.145</td>
<td>0.175</td>
<td>0.159</td>
</tr>
<tr>
<td>spam</td>
<td>0.826</td>
<td>0.056</td>
<td>0.839</td>
<td>0.826</td>
<td>0.832</td>
</tr>
</tbody>
</table>
Idea 3: “pure” hosts, incomplete records removed

<table>
<thead>
<tr>
<th>linguistic features</th>
<th>with</th>
<th>without</th>
</tr>
</thead>
<tbody>
<tr>
<td>instances</td>
<td>1211</td>
<td>1211</td>
</tr>
<tr>
<td>attributes</td>
<td>287</td>
<td>280</td>
</tr>
<tr>
<td>classified correctly</td>
<td>1099</td>
<td>1095</td>
</tr>
<tr>
<td>missclassified</td>
<td>112</td>
<td>116</td>
</tr>
</tbody>
</table>

- Further reduction of noisy examples results in quality improvement.
- The improvement gained from linguistic features is small, but clear.
## Idea 3: “pure” hosts, incomplete records removed

With linguistic features:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP</th>
<th>FP</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.970</td>
<td>0.089</td>
<td>0.961</td>
<td>0.970</td>
<td>0.966</td>
</tr>
<tr>
<td>undecided</td>
<td>0.306</td>
<td>0.031</td>
<td>0.294</td>
<td>0.306</td>
<td>0.300</td>
</tr>
<tr>
<td>spam</td>
<td>0.834</td>
<td>0.048</td>
<td>0.861</td>
<td>0.834</td>
<td>0.848</td>
</tr>
</tbody>
</table>

Without linguistic features:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP</th>
<th>FP</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.969</td>
<td>0.098</td>
<td>0.958</td>
<td>0.969</td>
<td>0.963</td>
</tr>
<tr>
<td>undecided</td>
<td>0.245</td>
<td>0.032</td>
<td>0.245</td>
<td>0.245</td>
<td>0.245</td>
</tr>
<tr>
<td>spam</td>
<td>0.834</td>
<td>0.048</td>
<td>0.861</td>
<td>0.834</td>
<td>0.848</td>
</tr>
</tbody>
</table>
All together: number of instances.
KYN – reference set, ALL – all records, PRUNED – without missing values, PURE – only pure hosts

All together: f-measure of the “spam” class.
Distribution of linguistic features in the Web-Spam2006UK corpus.

- Explore the distribution of each linguistic feature.
Distribution of linguistic features in the Web-Spam2006UK corpus.

- Explore the distribution of each linguistic feature.
- Explore fraction of spam within each range.
New results

Features
IDEA 1
IDEA 2
IDEA 3
Summary

percentage of pages

percentage of spam

lexical validity-[100-]
New results

Features

IDEA 1

IDEA 2

IDEA 3

Summary

percentage of pages

text-like fraction

percentage of spam

0% 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00

... of spam
text-like fraction
percantage of pages

0% 2% 4% 6% 8% 10% 12% 14% 16% 18%
New results

Features

IDEA 1
IDEA 2
IDEA 3
Summary

percentage of pages

percent of spam

passive voice
Conclusions

Preliminary experimental results seem to indicate:

- linguistic features introduced in [Zhou et al., 2004] slightly improve classification accuracy,
- pruning inconsistently labeled examples improves classification accuracy.

Further research:

- including other types of linguistic features (e.g. sentiment analysis, etc.),
- more systematic evaluation methods.
Acknowledgements

Ricardo Baeza-Yates\textsuperscript{Y,S}, Luca Becchetti\textsuperscript{R}, Paolo Boldi\textsuperscript{M}, Debora Donato\textsuperscript{Y}, Aristides Gionis\textsuperscript{Y}, Stefano Leonardi\textsuperscript{R}, Vanessa Murdock\textsuperscript{Y}, Massimo Santini\textsuperscript{M}, Fabrizio Silvestri\textsuperscript{P}, Sebastiano Vigna\textsuperscript{M}, Leszek Krupiński\textsuperscript{J}

Y. Yahoo! Research Barcelona – Catalunya, Spain
R. Università di Roma “La Sapienza” – Rome, Italy
S. Yahoo! Research Santiago – Chile
P. ISTI-CNR – Pisa, Italy
M. Università degli Studi di Milano – Milan, Italy
J. Polish-Japanese Institute of Information Technology, Poland
Thank you for your attention!


Combating Web spam with TrustRank.
In Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), pages 576–587, Toronto, Canada. Morgan Kaufmann.

Detecting spam web pages through content analysis.

ANF: a fast and scalable tool for data mining in massive graphs.

Automating Linguistics-Based Cues for Detecting Deception of Text-based Asynchronous Computer-Mediated Communication.