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

Part I

Introduction

Marcin Sydow

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)

Searching information – among the top Web activities

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:

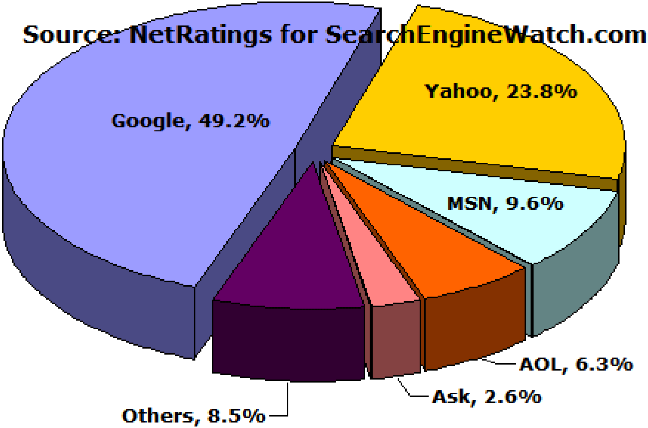
256.000.000 people used a search engine in December 2006
(Nielsen/NetRatings, 2006)

Some available statistics

500.000.000 queries per day globally (after Google, 2005)

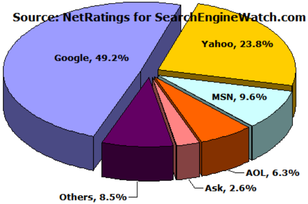
Some available statistics

500.000.000 queries per day globally (after Google, 2005)



Some available statistics

500.000.000 queries per day globally (after Google, 2005)

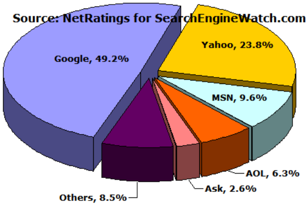


For a major global search engine it is:

- **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)

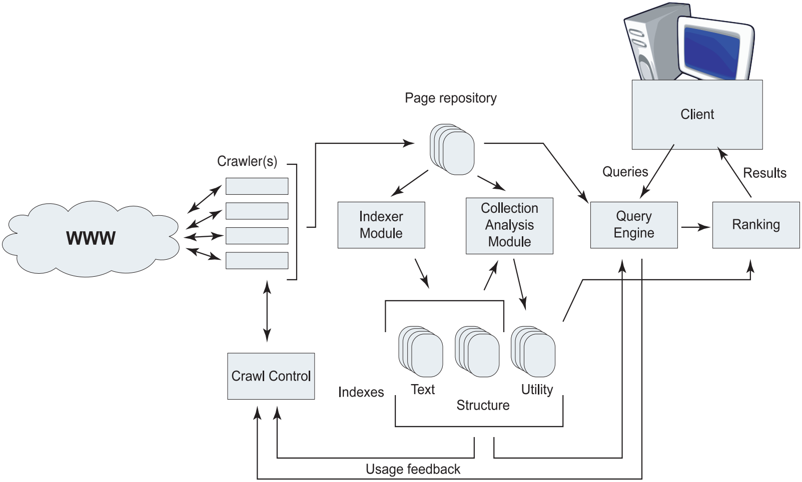


For a major global search engine it is:

- **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

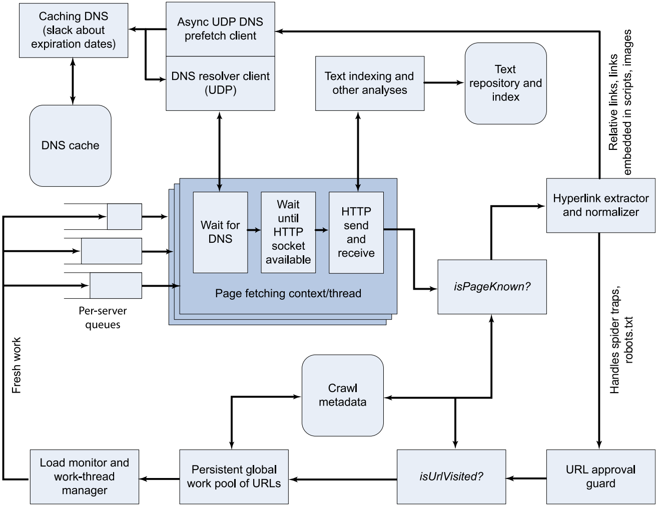
Search Engines – seemingly simple task

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

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

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

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 choosen out of **billions** other documents to link to (except link nepotism).

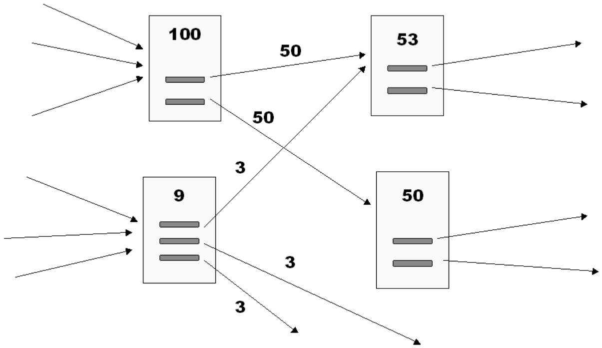
A simplistic assumption, but works in mass.

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



PageRank Equations

- simplified PageRank:

$$R(p) = \sum_{i \in \mathrm{IN}(p)} R(i)/\mathrm{outDeg}(i), \tag{1}$$

PageRank Equations

- simplified PageRank:

$$R(p) = \sum_{i \in \mathrm{IN}(p)} R(i)/\mathrm{outDeg}(i), \tag{1}$$

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

$$R(p) = (1-d) \sum_{i \in \mathrm{IN}(p)} \frac{R(i)}{\mathrm{outDeg}(i)} + d \cdot v(p) \tag{2}$$

PageRank Equations

- simplified PageRank:

$$R(p) = \sum_{i \in \mathrm{IN}(p)} R(i)/\mathrm{outDeg}(i), \tag{1}$$

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

$$R(p) = (1-d) \sum_{i \in \mathrm{IN}(p)} \frac{R(i)}{\mathrm{outDeg}(i)} + d \cdot v(p) \tag{2}$$

- simple “dangling-links” correction:

$$R(p) = (1-d) \sum_{i \in \mathrm{IN}(p)} \frac{R(i)}{\mathrm{outDeg}(i)} + d \cdot v(p) + (1-d) v(p) \sum_{i \in \mathrm{ZEROS}} R(i), \tag{3}$$

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)

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

(A. Broder, "Foundations of Web Advertising", tutorial, Edinburgh, 2006)

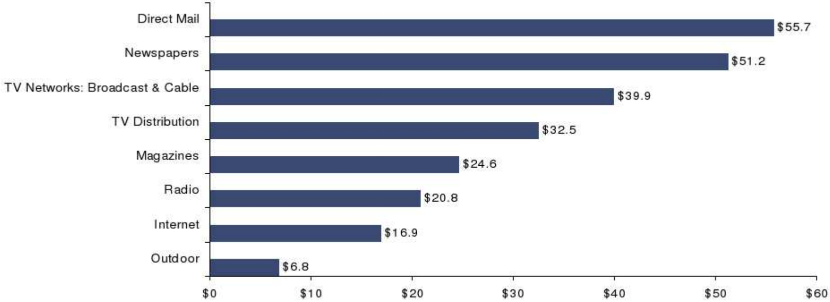
Main types:

- **sponsored links** (aside search results)
- **contextual ads** (placed on Web-sites)

Advertising Market Shares (USA, 2006)

- 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.

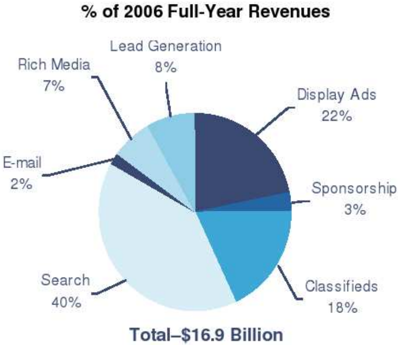
U.S. Advertising Market-Media Comparisons—2006 ($ Billions)

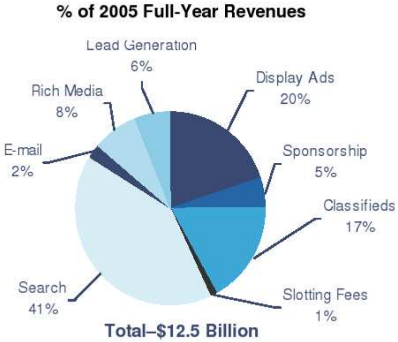


*The total U.S. advertising market is estimated at approximately $285 billion, and includes other segments not charted here.

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

The Central Role of Search Engines in WWW

Web pages are accessed through search engines

1. Search engine **ranking** $\longrightarrow$ Web page **visibility**
2. Web page **visibility** $\longrightarrow$ **traffic** on the page
3. **traffic** on the page $\longrightarrow$ **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)

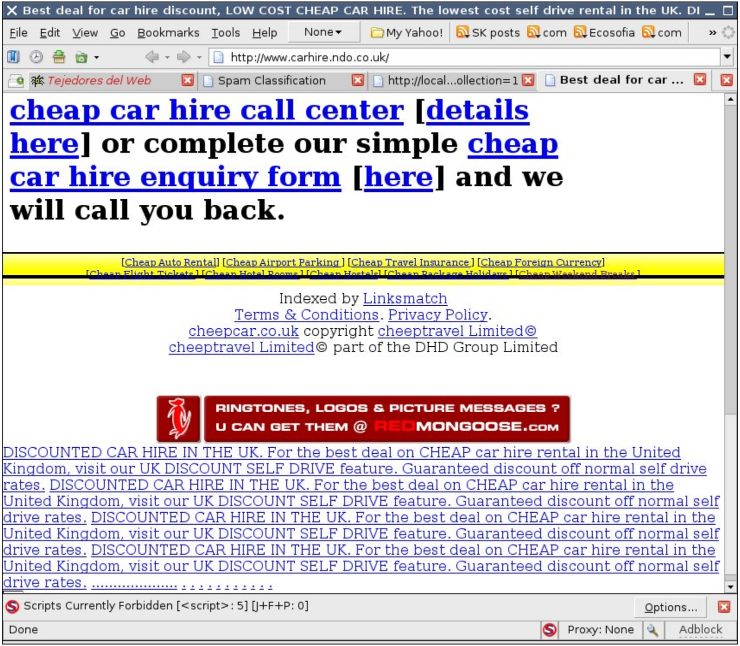
Spam techniques

- 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

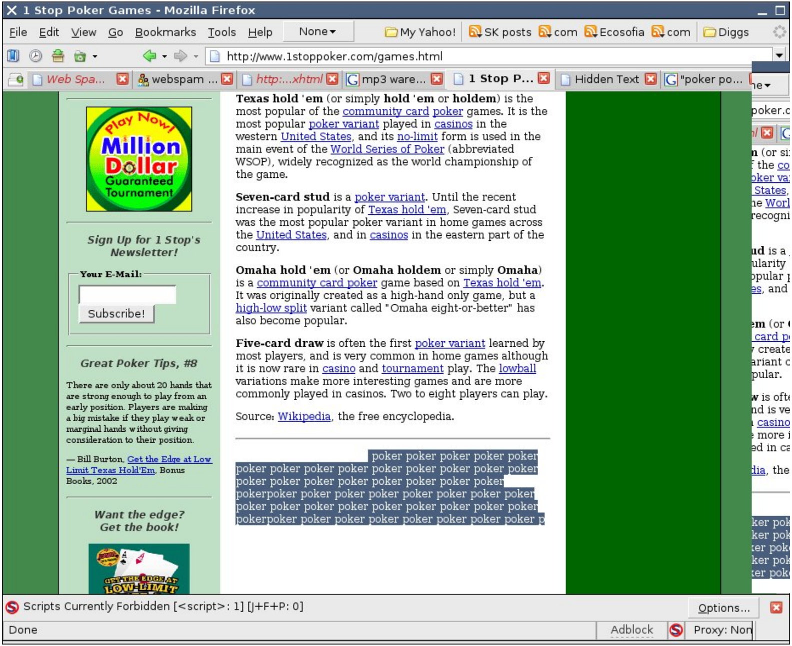
Spam techniques

- 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

Naïve Web Spam

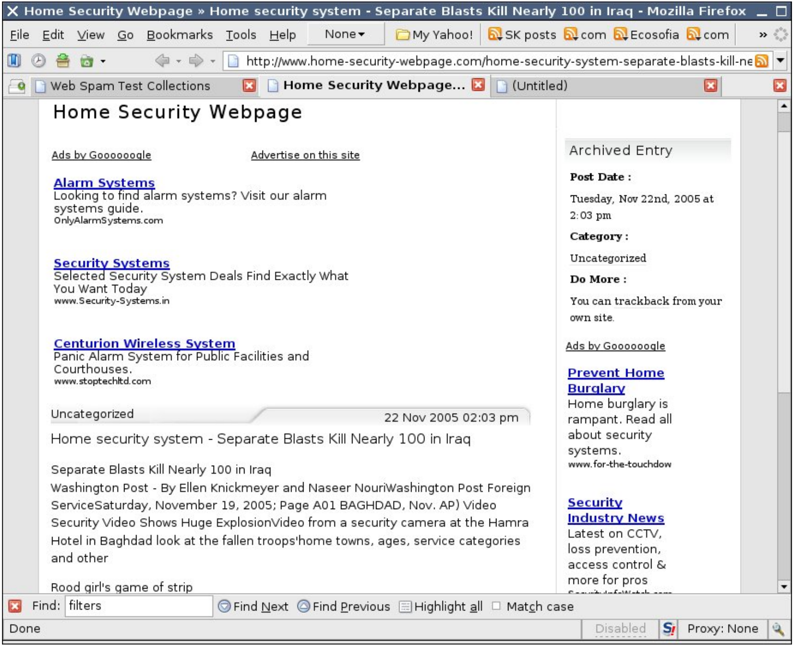


Hidden text

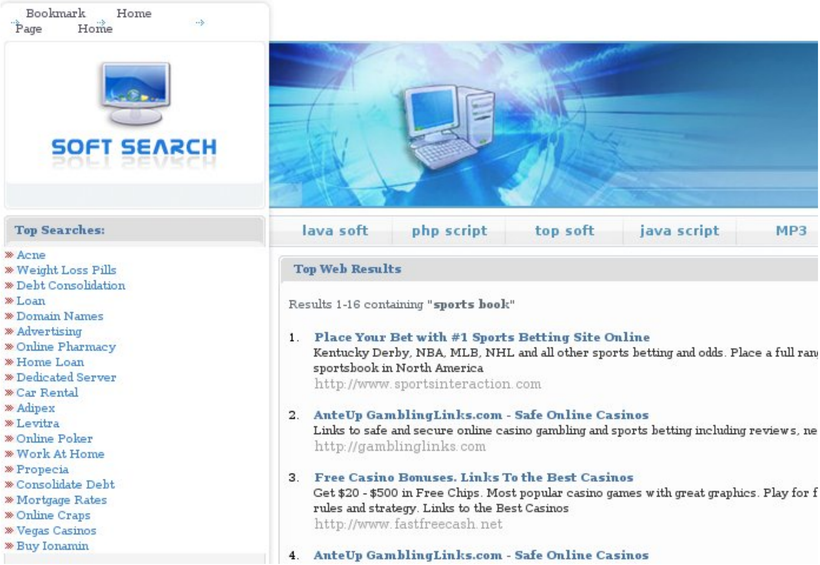


1 Stop Poker Games - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
None
My Yahoo! SK posts com Ecosofia com Diggs
http://www.1stoppoker.com/games.html
Web Spa... webspam ... http:...xhtml mp3 ware... 1 Stop P... Hidden Text "poker po...
Play Now! Million Dollar Guaranteed Tournament
Sign Up for 1 Stop's Newsletter!
Your E-Mail:
Subscribe!
Great Poker Tips, #8
There are only about 20 hands that are strong enough to play from an early position. Players are making a big mistake if they play weak or marginal hands without giving consideration to their position.
— Bill Burton, Get the Edge at Low Limit Texas Hold'Em, Bonus Books, 2002
Want the edge? Get the book!
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.
Source: Wikipedia, the free encyclopedia.
poker poker poker poker poker
poker poker poker poker poker poker poker poker poker
poker poker poker poker poker poker poker poker
pokerpoker poker poker poker poker poker poker poker
poker poker poker poker poker poker poker poker poker
pokerpoker poker poker poker poker poker poker poker p
Scripts Currently Forbidden [<script>: 1] [J+F+P: 0]
Options...
Done
Adblock
Proxy: Non

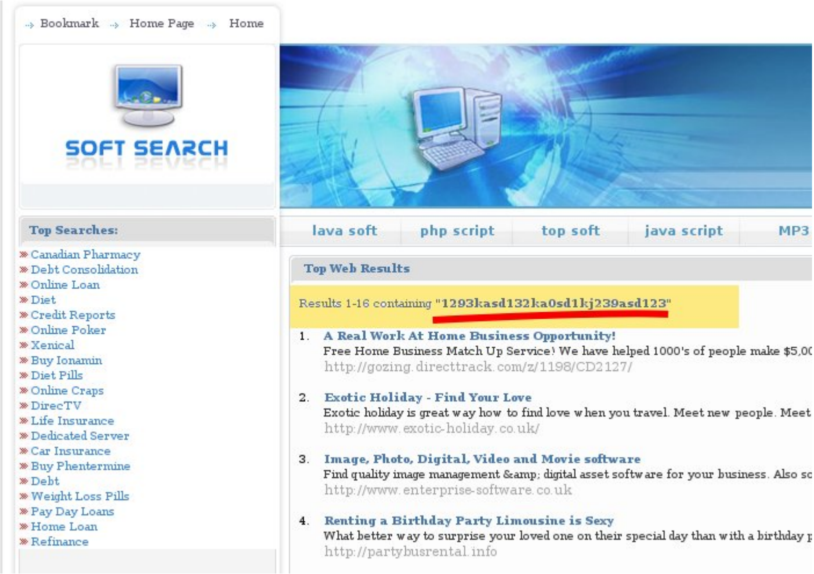
Made for Advertising



Search engine?



Fake search engine



"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.

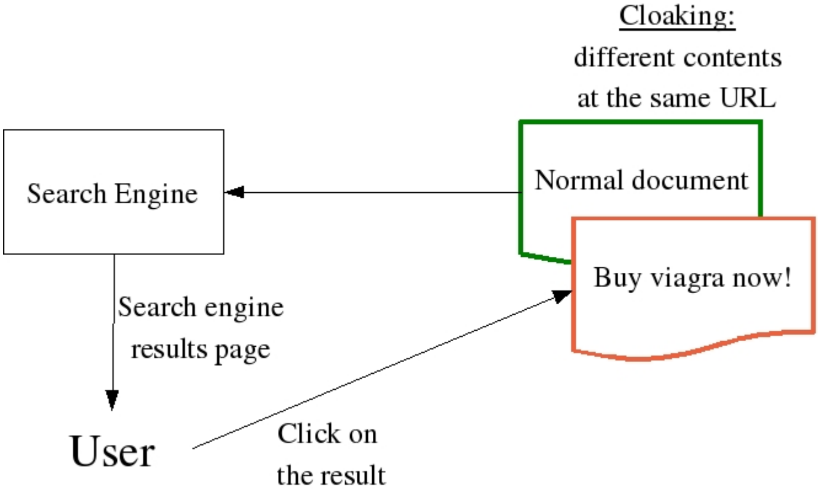
Look No further. You'll find it at Website design, management, marketing and promotion!

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

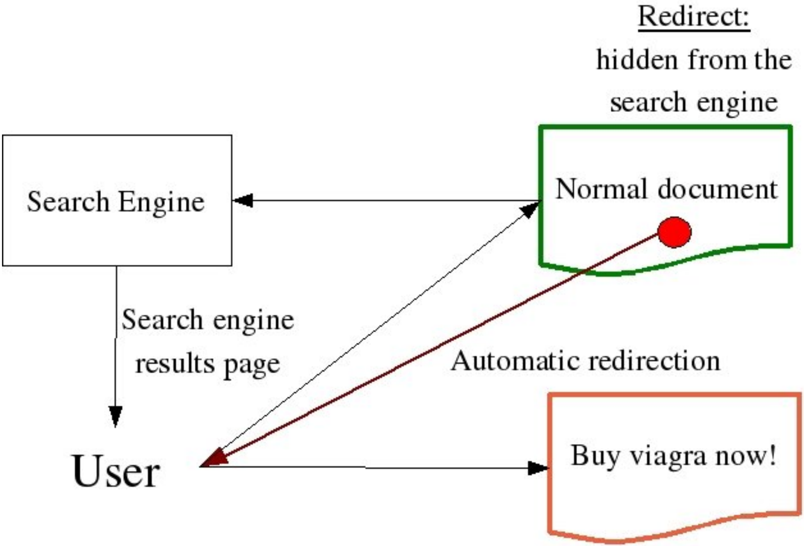
Website design, management, marketing and promotion brought to you by Dry Media, the leaders in this field.

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


Cloaking:
different contents
at the same URL
Search Engine
Normal document
Buy viagra now!
Search engine
results page
User
Click on
the result

Redirection



Redirects using Javascript

Simple redirect

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

“Hidden” redirect

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

Problem: obfuscated code

Obfuscated redirect

```
<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

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

More examples: [Chellapilla and Maykov, 2007]

Fighting Spam

On the search engines' side:

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)

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”**:
  1. spammers apply new deceptive technique
  2. 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...

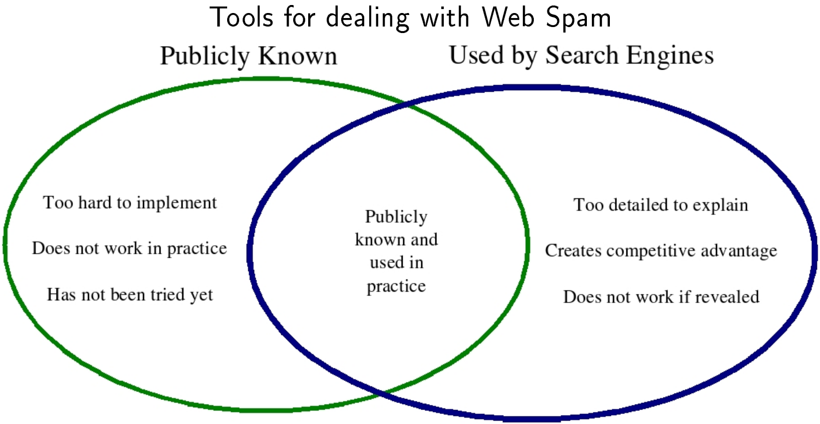
Machine Learning approach recently applied to support Search Engines in combating Web spam

Part II

Reference Corpus & State of the Art

Carlos Castillo

Tools for dealing with Web Spam



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

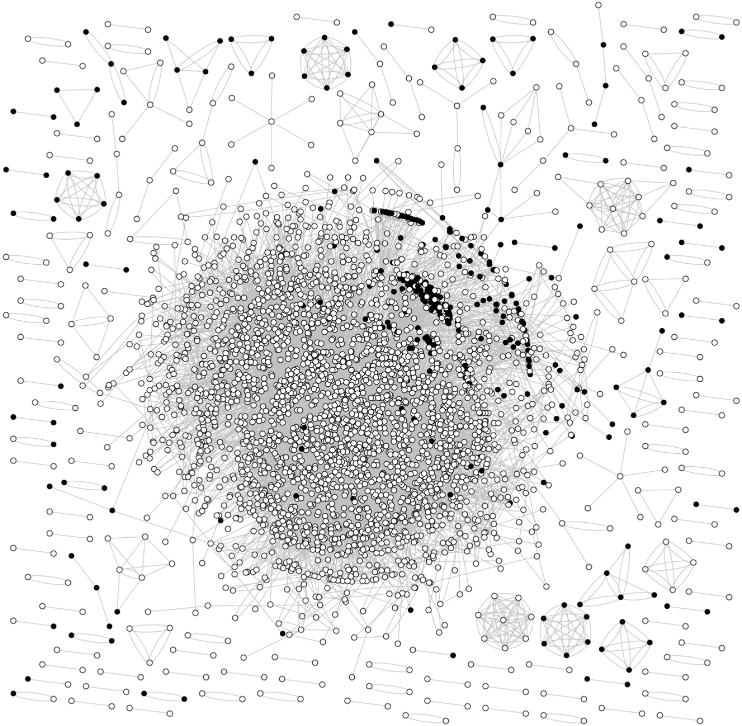
Machine Learning Challenges:

- 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



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)

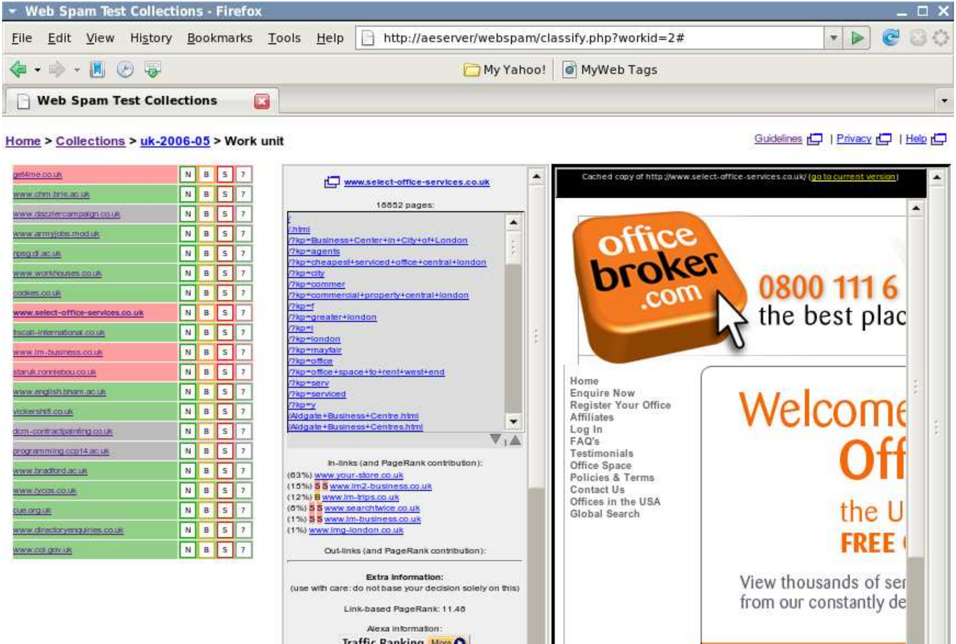
Crawling of base data

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 $<=$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

Labeling process

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

Agreement

2547	infoserve.qub.ac.uk	AUTO_domain:N		
2548	inhale.hud.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2549	intoanpaydaytexas.dahannusaprima.co.uk	josiane:S	brian:S	
2550	insight.zdnet.co.uk	AUTO_dmoz:N	wenzhong:N	
2552	instruments.abcaz.co.uk	antonio:N	chato:B	
2554	insurance.onlineholidays.org.uk	xiaoguang:N	tamas:S	chato:S
2557	intarch.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2558	interact.brighton.ac.uk	AUTO_domain:N		
2559	internal.bath.ac.uk	AUTO_domain:N		
2560	internal.cs.ncl.ac.uk	AUTO_domain:N		
2561	internal.iop.kcl.ac.uk	AUTO_domain:N		
2564	internetmegastores.co.uk	thomas:S	antonio:N	chato:B
2565	internt.nhm.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2566	intranet.bton.ac.uk:259	antonio:?	chato:N	
2567	intranet.ee.ic.ac.uk	AUTO_domain:N		
2568	intranet.londonmet.ac.uk	AUTO_domain:N		
2569	intranet.open.ac.uk	AUTO_domain:N		
2570	intranet.salford.ac.uk	AUTO_domain:N		
2571	investing.reuters.co.uk	thomas:N	mike:N	
2572	investing.thisismoney.co.uk	miral:N	alex:B	
2573	ioewebserver.ioe.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2574	iqi.iee.org.uk	omar:N	zoltan:N	
2575	ion.le.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2577	ipac.brad.ac.uk	AUTO_domain:N		
2578	ipod.skdirectory.co.uk	miral:S	luca:?	
2579	iq.lycos.co.uk	pascal:N	zoltan:N	
2580	ir.dcs.gla.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2581	irish-swingers.connect4fun.co.uk	tanguy:S	zoltan:S	
2583	isdocs.brighton.ac.uk	AUTO_domain:N		
2584	it.auctionsista.co.uk	zoltan:S	sebastiano:S	
2585	itreviews.co.uk	massimo:N	brian:N	
2586	itweek.directorym.co.uk	xiaoguang:N	thiago:N	

Results

- Labels:

Label	Frequency	Percentage
normal	4,046	61.75%
borderline	709	10.82%
spam	1,447	22.08%
can not classify	350	5.34%

- Agreement:

Category	Kappa	Interpretation
normal	0.62	Substantial agreement
spam	0.63	Substantial agreement
borderline	0.11	Slight agreement
global	0.56	Moderate agreement

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/**

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/**
- 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 (`.ac.uk`, `.sch.uk`, `.gov.uk`, `.mod.uk`, `.nhs.uk` or `.police.uk`)
  - **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/**
- 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/**
- Web Spam challenge
  - Track I: Information retrieval + Machine learning
  - Track II: Machine learning
  - **http://webspam.lip6.fr/**

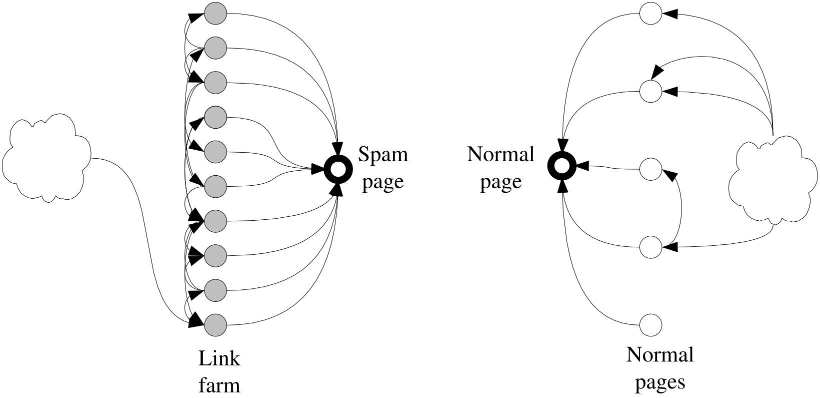
Result: first public Web Spam collection

- 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/**
- Web Spam challenge
  - Track I: Information retrieval + Machine learning
  - Track II: Machine learning
  - **http://webspam.lip6.fr/**
- AIRWeb 2007 Workshop

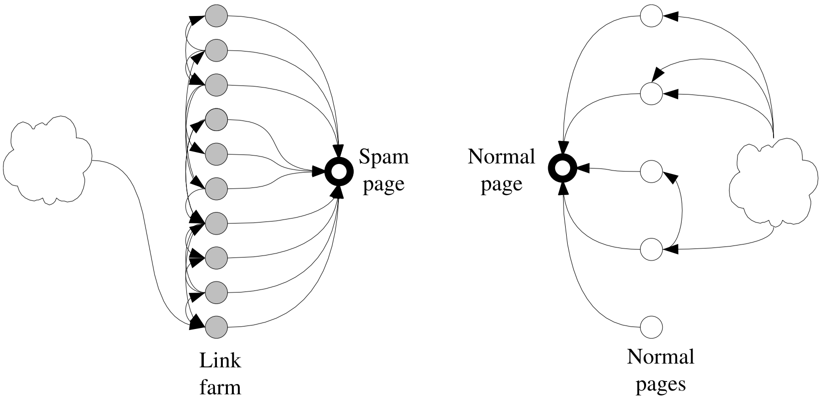
Result: first public Web Spam collection

- 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/**
- Web Spam challenge
  - Track I: Information retrieval + Machine learning
  - Track II: Machine learning
  - **http://webspam.lip6.fr/**
- AIRWeb 2007 Workshop
- GraphLab 2007 Workshop

Link farms



Link farms



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

Semi-streaming model

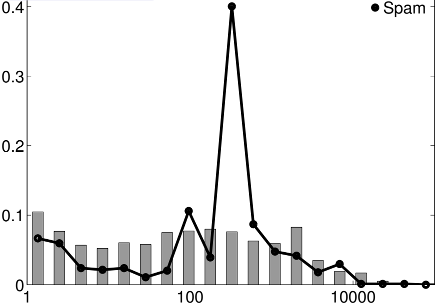
Handling large graphs:

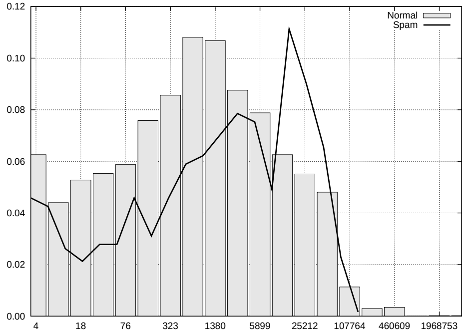
- Memory size enough to hold some data per-node
- Disk size enough to hold some data per-edge
- A small number of sequential passes over the data

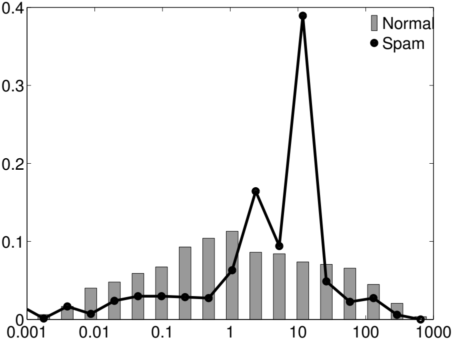
Link-Based Features

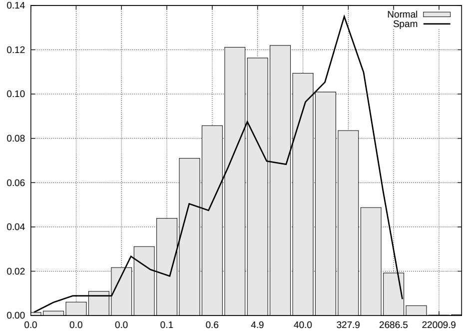
- Degree-related measures
- PageRank
- TrustRank [Gyöngyi et al., 2004]
- Truncated PageRank [Becchetti et al., 2006]
- Estimation of supporters [Becchetti et al., 2006]

Degree-Based

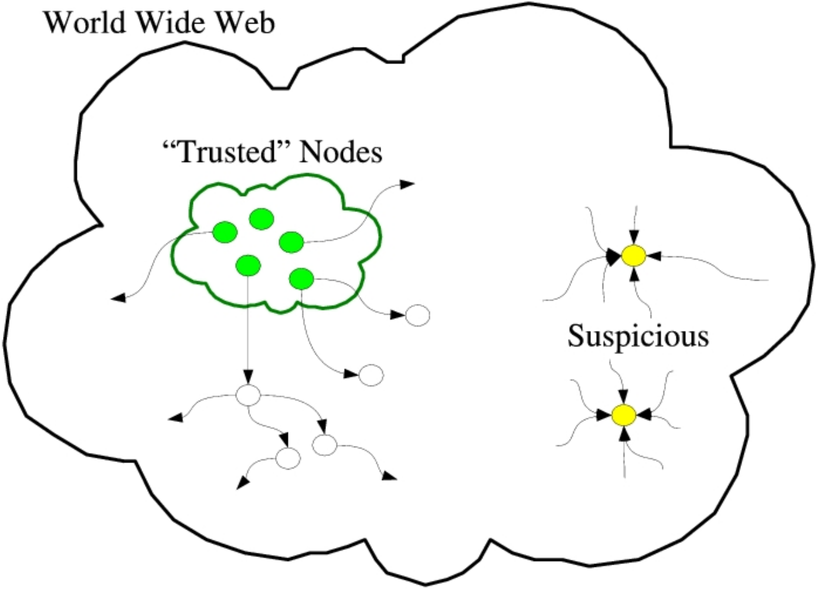


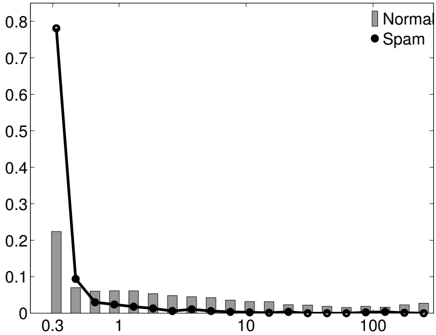


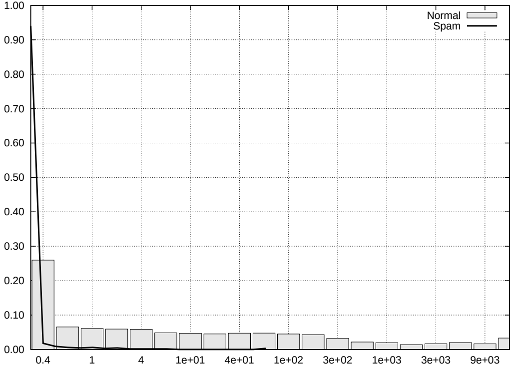




TrustRank Idea

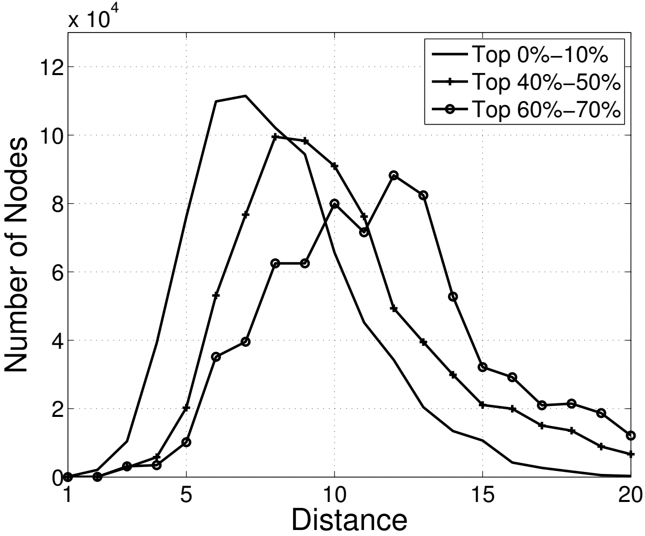




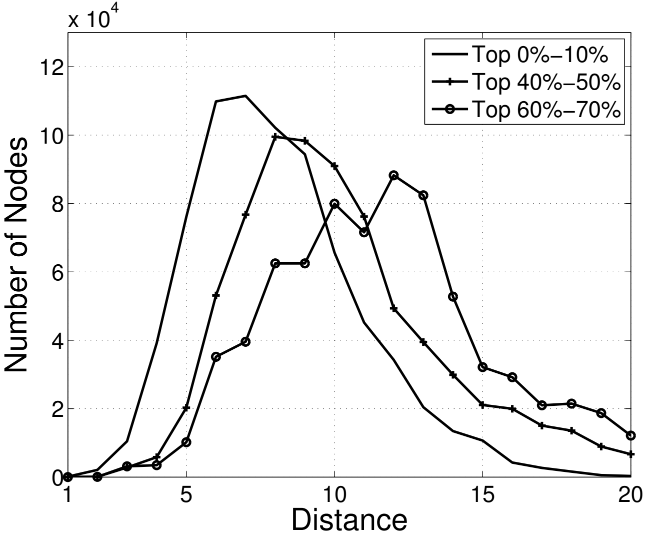


TrustRank / PageRank

Hop-plot and PageRank

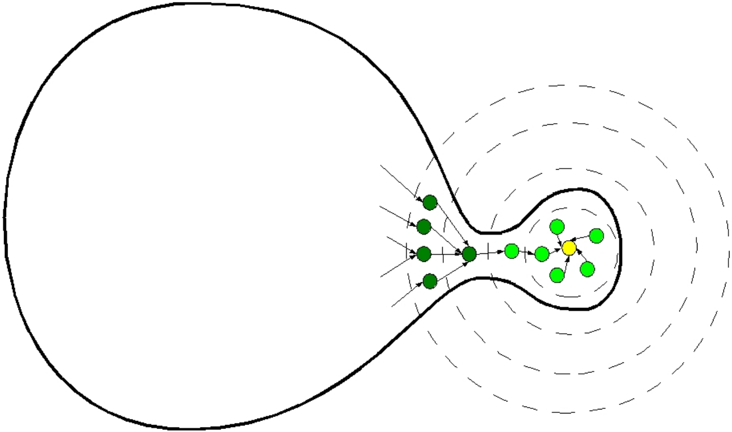


Hop-plot and PageRank

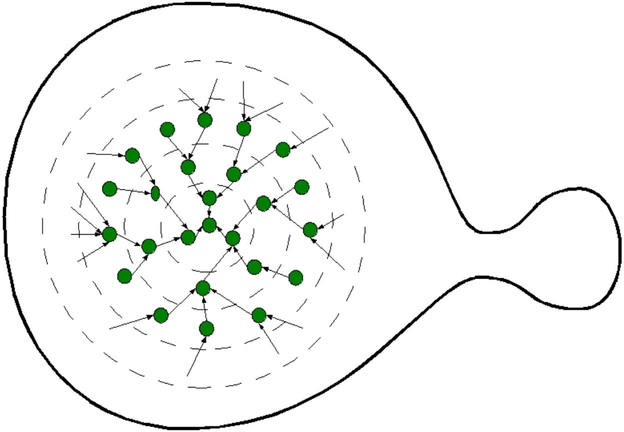


Areas below the curves are equal if we are in the same strongly-connected component

Neighbors: spam

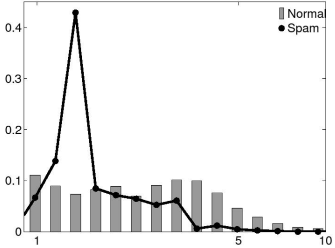


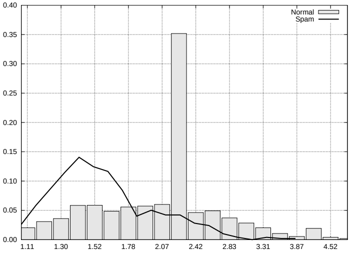
Neighbors: normal



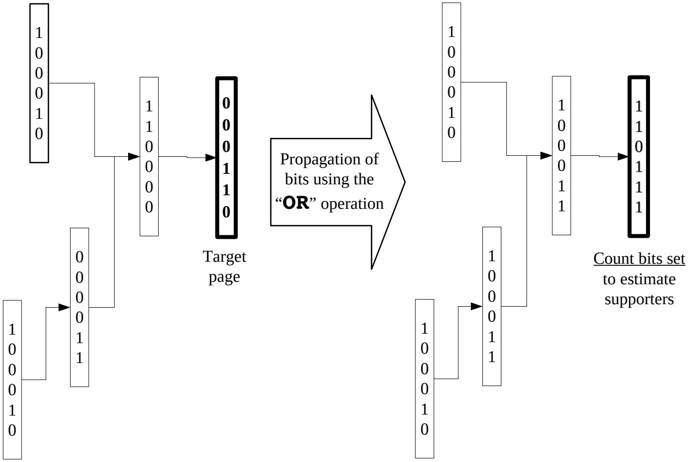
Bottleneck number

$b_d(x) = \min_{j \le d} \{|N_j(x)|/|N_{j-1}(x)|\}$. 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.

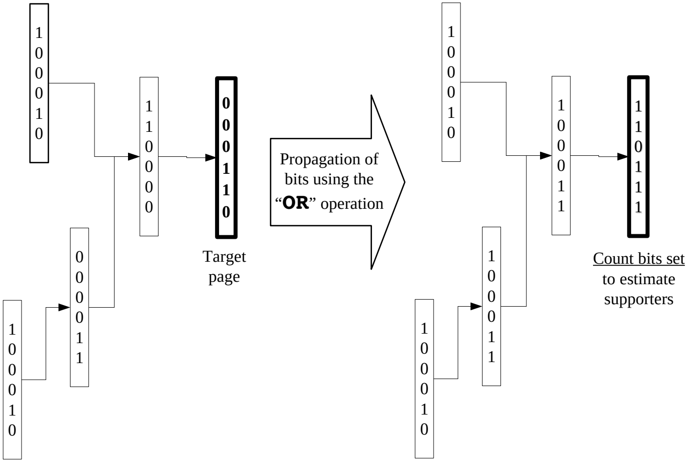




Probabilistic counting



Probabilistic counting



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

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" $|P \cap F|/|P|$
- Corpus "recall" $|P \cap F|/|F|$
- Query "precision" $|P \cap Q|/|P|$
- Query "recall" $|P \cap Q|/|Q|$

Average word length

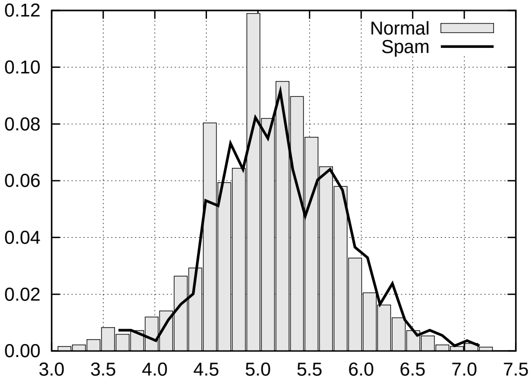


Figure: Histogram of the average word length in non-spam vs. spam pages for $k = 500$.

Corpus precision

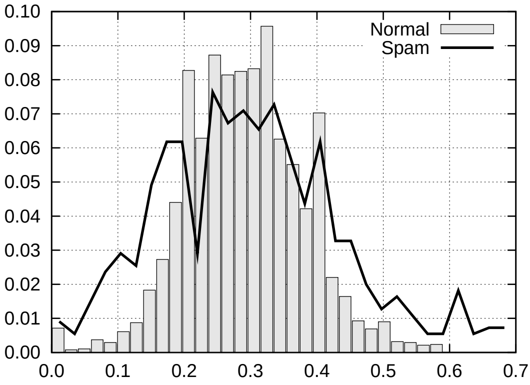


Figure: Histogram of the corpus precision in non-spam vs. spam pages.

Query precision

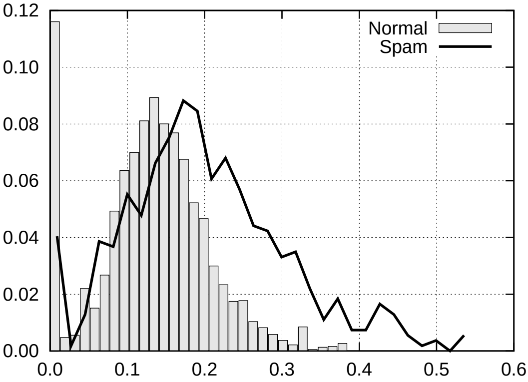


Figure: Histogram of the query precision in non-spam vs. spam pages for $k = 500$.

Cost-sensitive decision tree with bagging

Bagging of 10 decision trees, asymmetrical costs.

Cost ratio	1	10	20	30	50
True positive rate	65.8%	66.7%	71.1%	78.7%	84.1%
False positive rate	2.8%	3.4%	4.5%	5.7%	8.6%
F-Measure	0.712	0.703	0.704	**0.723**	0.692

Link- and content-based features

Link-based and content-based

	Both	Link-only	Content-only
True positive rate	78.7%	79.4%	64.9%
False positive rate	5.7%	9.0%	3.7%
F-Measure	**0.723**	0.659	0.683

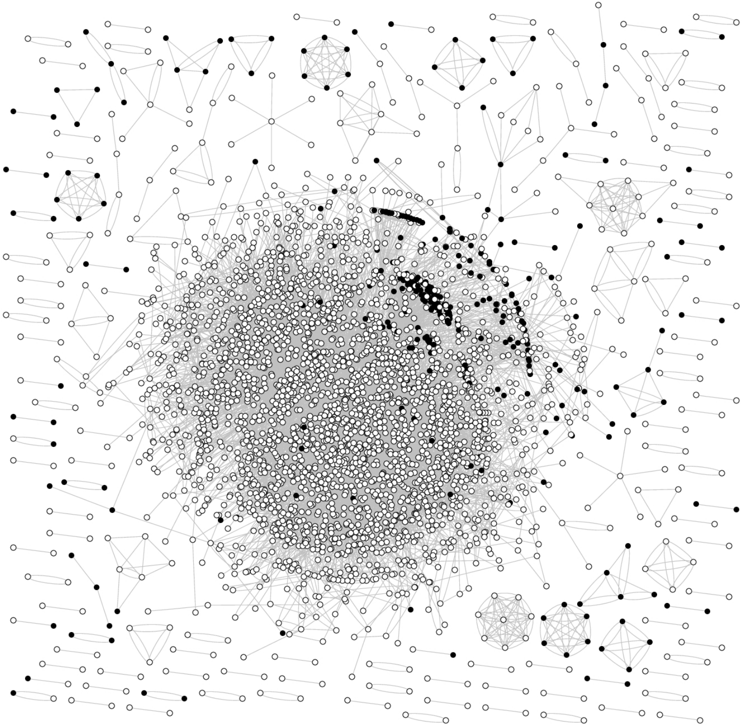
General hypothesis

Pages topologically close to each other are more likely to have the same label (spam/nonspam) 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.

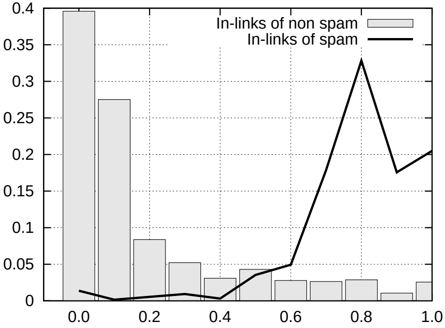
Pages linked together are more likely to be on the same topic than random pairs of pages [Davison, 2000]



Topological dependencies: in-links

Histogram of fraction of spam hosts in the in-links

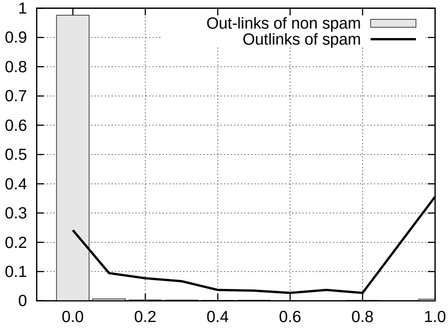
- 0 = no in-link comes from spam hosts
- 1 = all of the in-links come from spam hosts



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

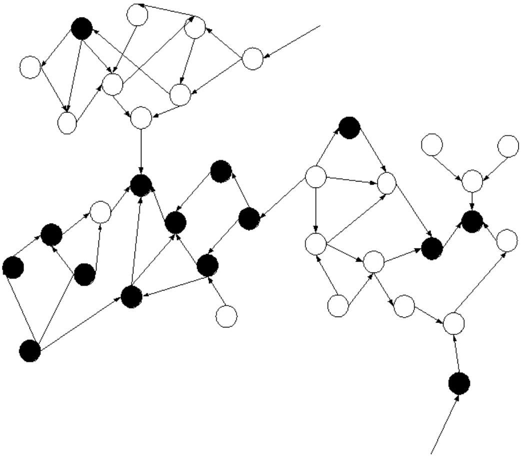


Idea 1: Clustering

Classify, then cluster hosts, then assign the same label to all hosts in the same cluster by majority voting

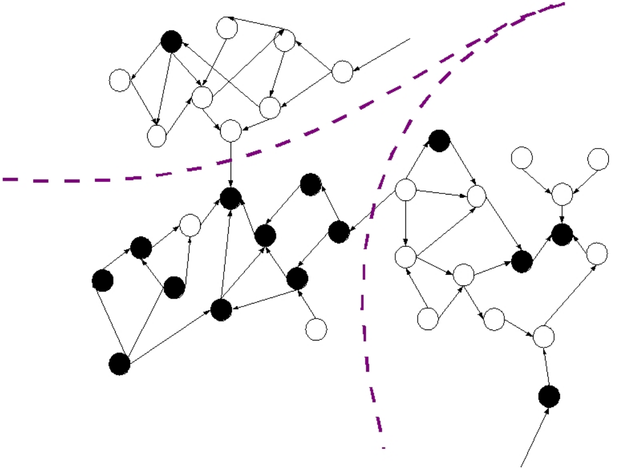
Idea 1: Clustering (cont.)

Initial prediction:



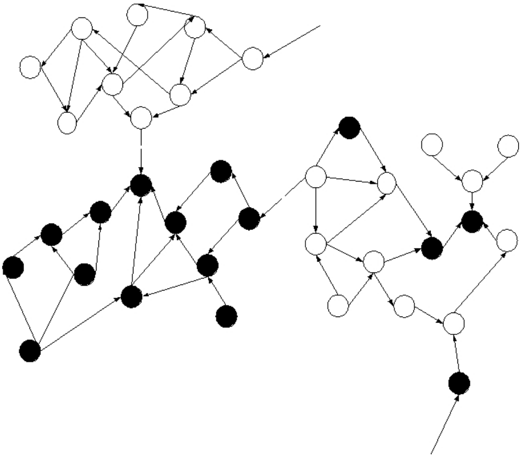
Idea 1: Clustering (cont.)

Clustering:



Idea 1: Clustering (cont.)

Final prediction:



Idea 1: Clustering – Results

	Baseline	Clustering
Without bagging		
True positive rate	75.6%	74.5%
False positive rate	8.5%	6.8%
F-Measure	0.646	**0.673**
With bagging		
True positive rate	78.7%	76.9%
False positive rate	5.7%	5.0%
F-Measure	0.723	0.728

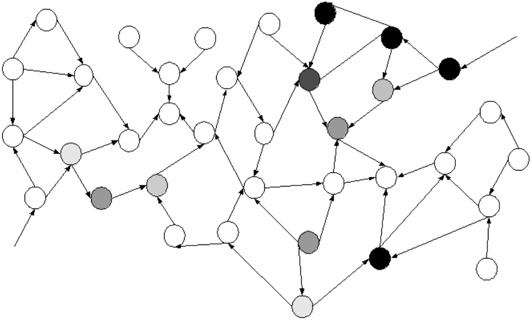
☑ Reduces error rate

Idea 2: Propagate the label

Classify, then interpret "spamicity" as a probability, then do a random walk with restart from those nodes

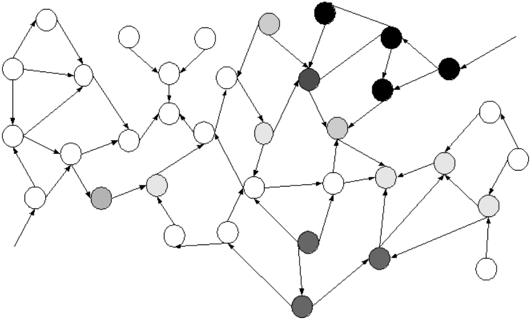
Idea 2: Propagate the label (cont.)

Initial prediction:



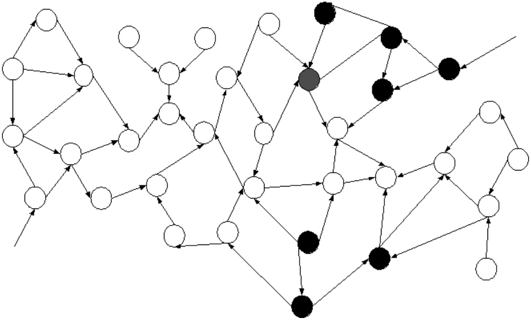
Idea 2: Propagate the label (cont.)

Propagation:



Idea 2: Propagate the label (cont.)

Final prediction, applying a threshold:



Idea 2: Propagate the label – Results

	Baseline	Fwds.	Backwds.	Both
Classifier without bagging				
True positive rate	75.6%	70.9%	69.4%	71.4%
False positive rate	8.5%	6.1%	5.8%	5.8%
F-Measure	0.646	0.665	0.664	**0.676**
Classifier with bagging				
True positive rate	78.7%	76.5%	75.0%	75.2%
False positive rate	5.7%	5.4%	4.3%	4.7%
F-Measure	0.723	0.716	0.733	0.724

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.)

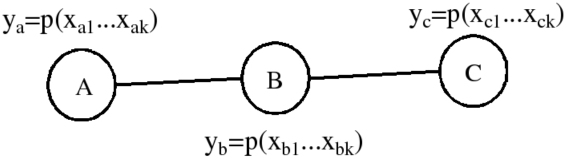
- Let $p(x) \in [0..1]$ be the prediction of a classification algorithm for a host x using k features
- Let $N(x)$ be the set of pages related to x (in some way)
- Compute

$$f(x) = \frac{\sum_{g \in N(x)} p(g)}{|N(x)|}$$

- Add $f(x)$ as an extra feature for instance x and learn a new model with $k+1$ features

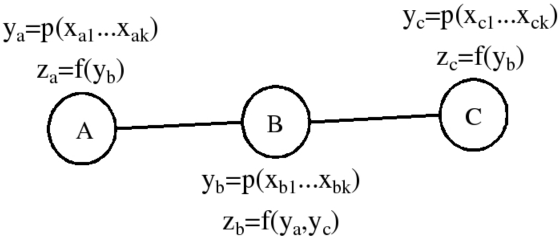
Idea 3: Stacked graphical learning (cont.)

Initial prediction:



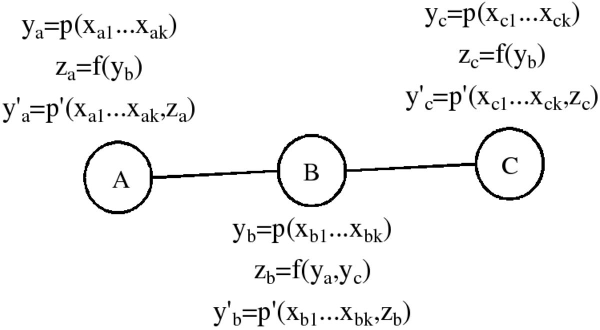
Idea 3: Stacked graphical learning (cont.)

Computation of new feature:



Idea 3: Stacked graphical learning (cont.)

New prediction with $k+1$ features:



Idea 3: Stacked graphical learning - Results

	Baseline	Avg. of in	Avg. of out	Avg. of both
True positive rate	78.7%	84.4%	78.3%	85.2%
False positive rate	5.7%	6.7%	4.8%	6.1%
F-Measure	0.723	0.733	0.742	**0.750**

☑ Increases detection rate

Idea 3: Stacked graphical learning x2

And repeat ...

	Baseline	First pass	Second pass
True positive rate	78.7%	85.2%	88.4%
False positive rate	5.7%	6.1%	6.3%
F-Measure	0.723	0.750	**0.763**

☑ Significant improvement over the baseline

Part III

New Experimental Results

Jakub Piskorski, Marcin Sydow, Dawid Weiss

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 exstensively 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

	linguistic features			
	with		without	
instances	8 411		8 411	
attributes	287		280	
classiffied correctly	7 666	91.14%	7 687	91.39%
missclassified	745	8.85%	724	8.60%

- The results are not much different.

Idea 1: Just linguistic features

Figures in red are “better”.

With linguistic features:

Class	TP	FP	Precision	Recall	F-Measure
normal	0.970	0.435	0.946	0.970	0.958
undecided	0.091	0.010	0.162	0.091	0.116
spam	0.525	0.033	0.615	0.525	0.566

Without linguistic features:

Class	TP	FP	Precision	Recall	F-Measure
normal	0.970	0.415	0.949	0.970	0.959
undecided	0.108	0.010	0.186	0.108	0.137
spam	0.552	0.033	0.629	0.552	0.588

IDEA 2

Prune the input by removing records with missing values.
Rerun the experiments with and without linguistic attributes.

Idea 2: prune records with missing values

	linguistic features			
	with		without	
instances	6 644		6 644	
attributes	287		280	
classiffied correctly	6 016	90.54%	6 009	90.44%
missclassified	628	9.45%	635	9.55%

- Not much improvement (difference so small it is most likely statistically insignificant).

Idea 2: prune records with missing values

With linguistic features:

Class	TP	FP	Precision	Recall	F-Measure
normal	0.958	0.343	0.954	0.958	0.956
undecided	0.112	0.019	0.119	0.112	0.115
spam	0.608	0.039	0.622	0.608	0.615

Without linguistic features:

Class	TP	FP	Precision	Recall	F-Measure
normal	0.958	0.348	0.954	0.958	0.956
undecided	0.105	0.019	0.113	0.105	0.109
spam	0.601	0.039	0.616	0.601	0.608

IDEA 3

Choose only “pure” hosts (for which class decision was univocal).
Rerun the experiments with and without linguistic attributes.

Pure hosts – explanation

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;

- 1 049 NNN hosts,
- 391 SS hosts,
- 57 BB hosts,
- (no SSS or BBB examples in the original data).

The above gives a total of 1 497 pure hosts used as input.

Idea 3: "pure" hosts

	linguistic features			
	with		without	
instances	1 497		1 497	
attributes	287		280	
classiffied correctly	1 328	88.71%	1 330	88.84%
missclassified	169	11.28%	167	11.15%

Idea 3: “pure” hosts

With linguistic features:

Class	TP	FP	Precision	Recall	F-Measure
normal	0.949	0.107	0.954	0.949	0.952
undecided	0.193	0.042	0.155	0.193	0.172
spam	0.821	0.055	0.840	0.821	0.831

Without linguistic features:

Class	TP	FP	Precision	Recall	F-Measure
normal	0.950	0.103	0.956	0.950	0.953
undecided	0.175	0.041	0.145	0.175	0.159
spam	0.826	0.056	0.839	0.826	0.832

Idea 3: "pure" hosts, incomplete records removed

	linguistic features			
	with		without	
instances	1 211		1 211	
attributes	287		280	
classiffied correctly	1 099	90.75%	1 095	90.42%
missclassified	112	9.24%	116	9.57%

- Further reduction of noisy examples results in quality improvement.
- The improvement gained from linguistic features is small, but clear.

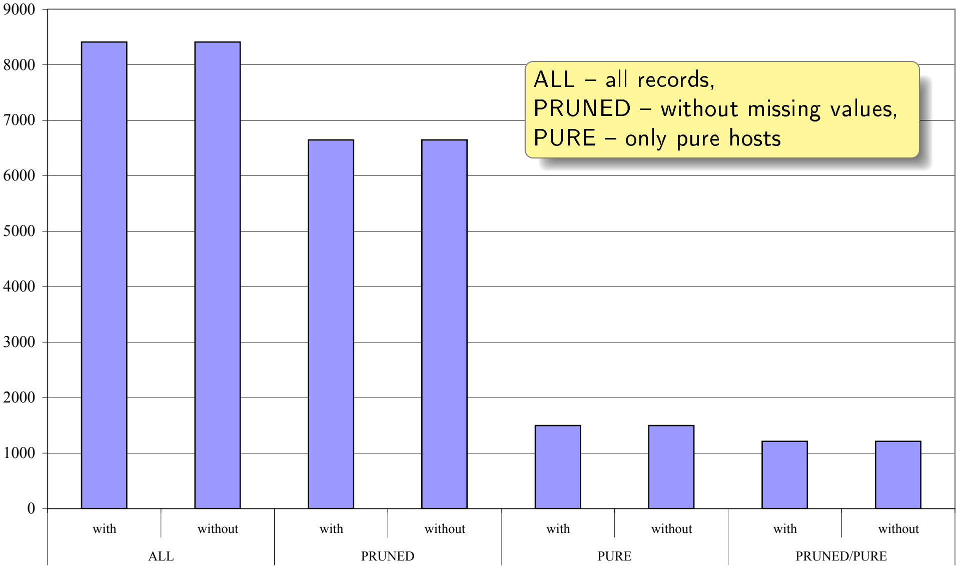
Idea 3: "pure" hosts, incomplete records removed

With linguistic features:

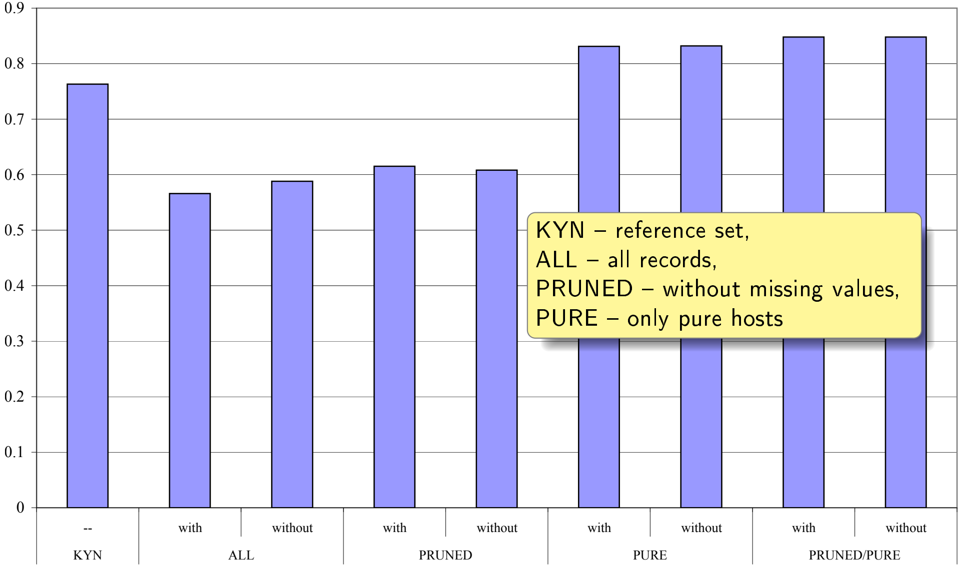
Class	TP	FP	Precision	Recall	F-Measure
normal	0.970	0.089	0.961	0.970	0.966
undecided	0.306	0.031	0.294	0.306	0.300
spam	0.834	0.048	0.861	0.834	0.848

Without linguistic features:

Class	TP	FP	Precision	Recall	F-Measure
normal	0.969	0.098	0.958	0.969	0.963
undecided	0.245	0.032	0.245	0.245	0.245
spam	0.834	0.048	0.861	0.834	0.848



All together: number of instances.



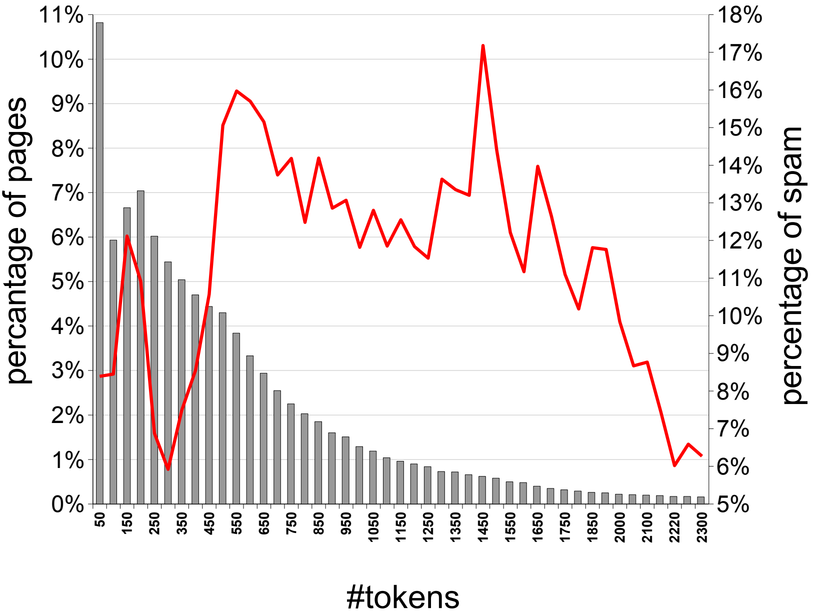
All together: f-measure of the “spam” class.

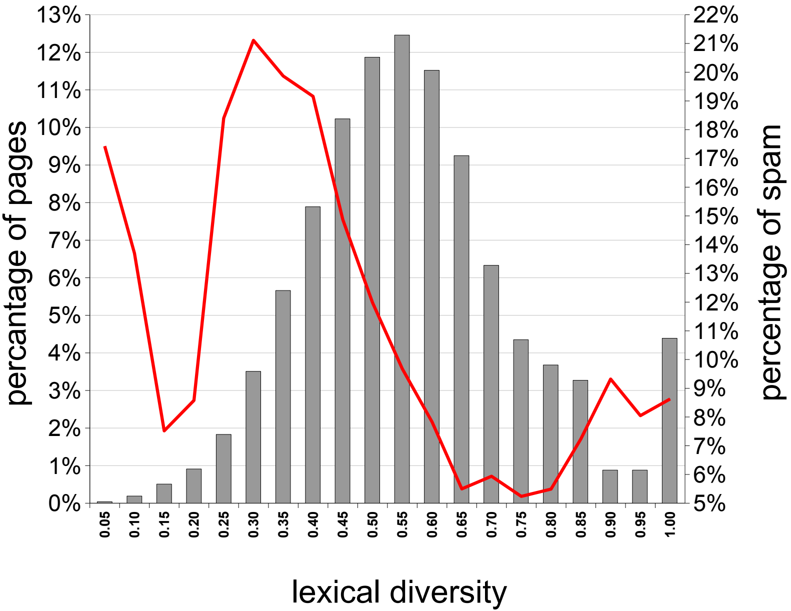
Distribution of linguistic features in the Web-Spam2006UK corpus.

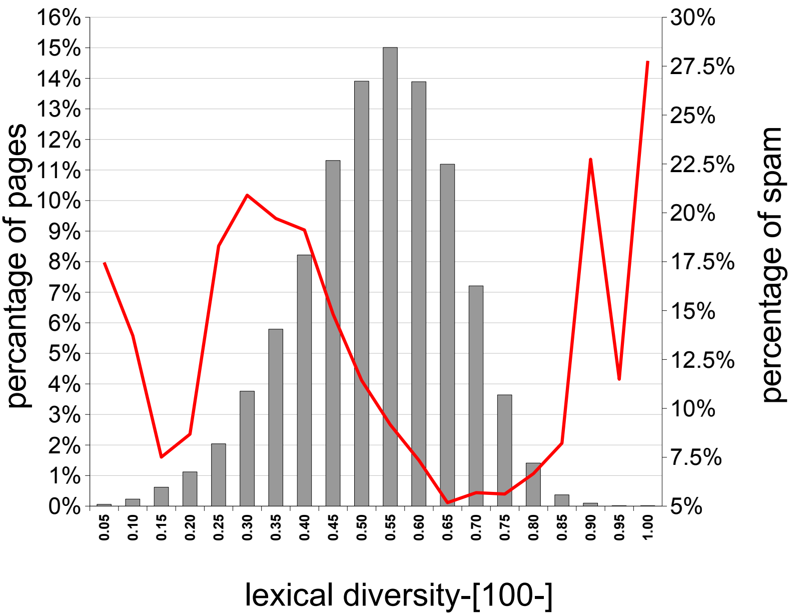
- Explore the distribution of each linguistic feature.

Distribution of linguistic features in the Web-Spam2006UK corpus.

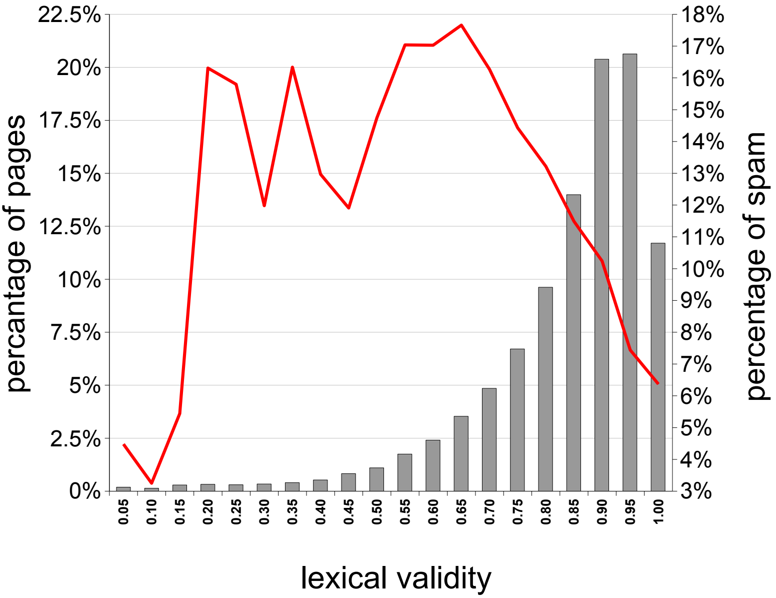
- Explore the distribution of each linguistic feature.
- Explore fraction of spam within each range.

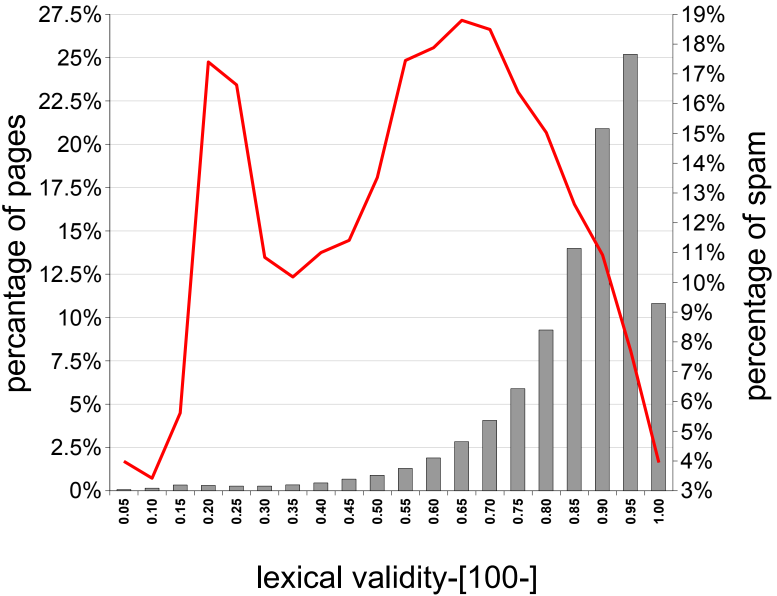

percantage of pages
percentage of spam
#tokens
11%
10%
9%
8%
7%
6%
5%
4%
3%
2%
1%
0%
18%
17%
16%
15%
14%
13%
12%
11%
10%
9%
8%
7%
6%
5%
50
150
250
350
450
550
650
750
850
950
1050
1150
1250
1350
1450
1550
1650
1750
1850
2000
2100
2220
2300

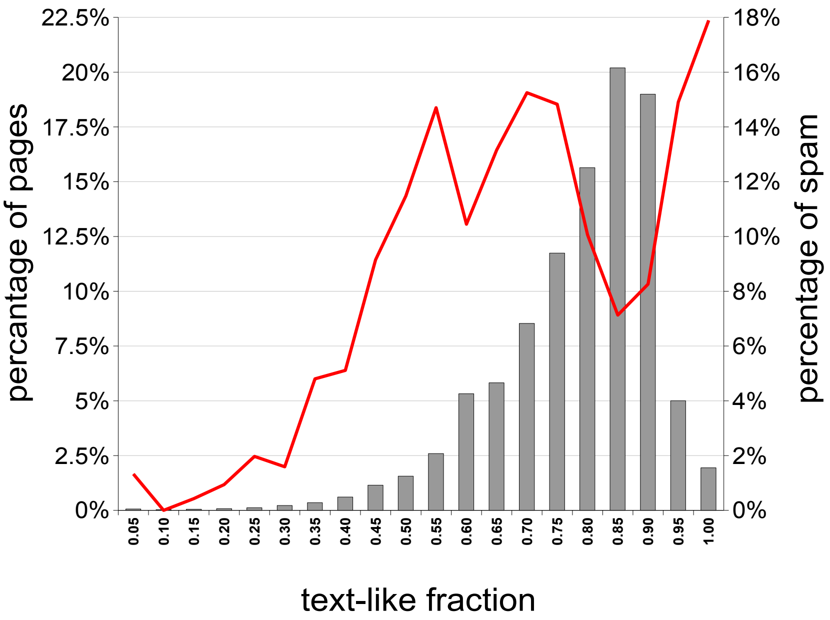

percantage of pages
percentage of spam
13%
12%
11%
10%
9%
8%
7%
6%
5%
4%
3%
2%
1%
0%
22%
21%
20%
19%
18%
17%
16%
15%
14%
13%
12%
11%
10%
9%
8%
7%
6%
5%
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
lexical diversity

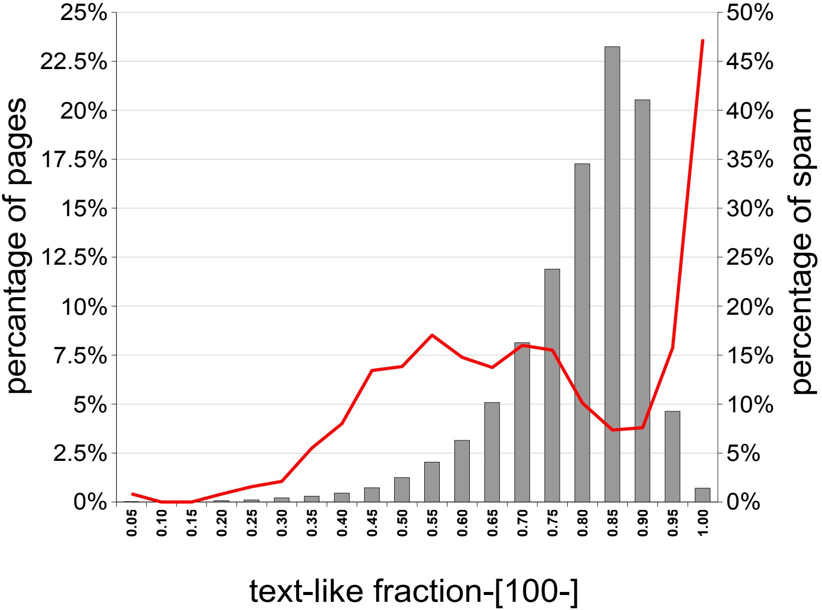


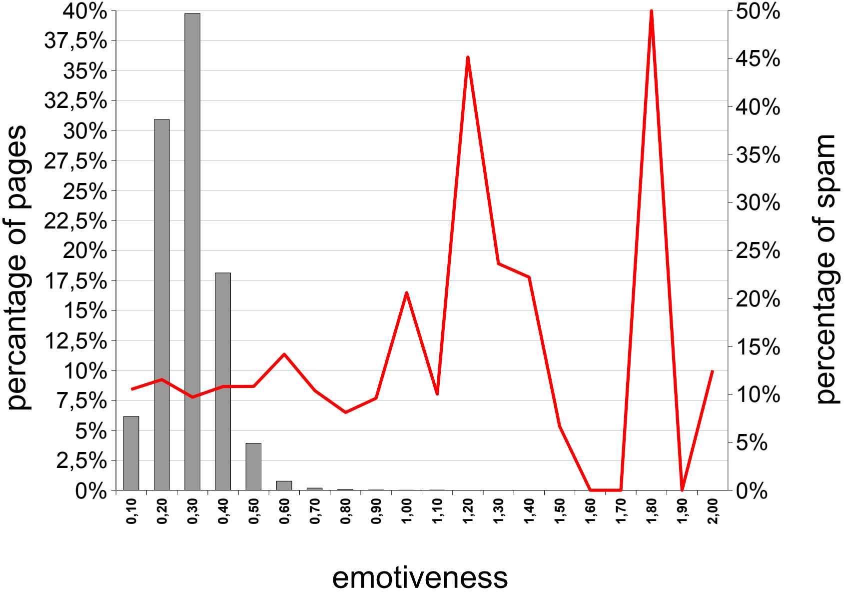
16%
15%
14%
13%
12%
11%
10%
9%
8%
7%
6%
5%
4%
3%
2%
1%
0%
percantage of pages
30%
27.5%
25%
22.5%
20%
17.5%
15%
12.5%
10%
7.5%
5%
percentage of spam
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
lexical diversity-[100-]

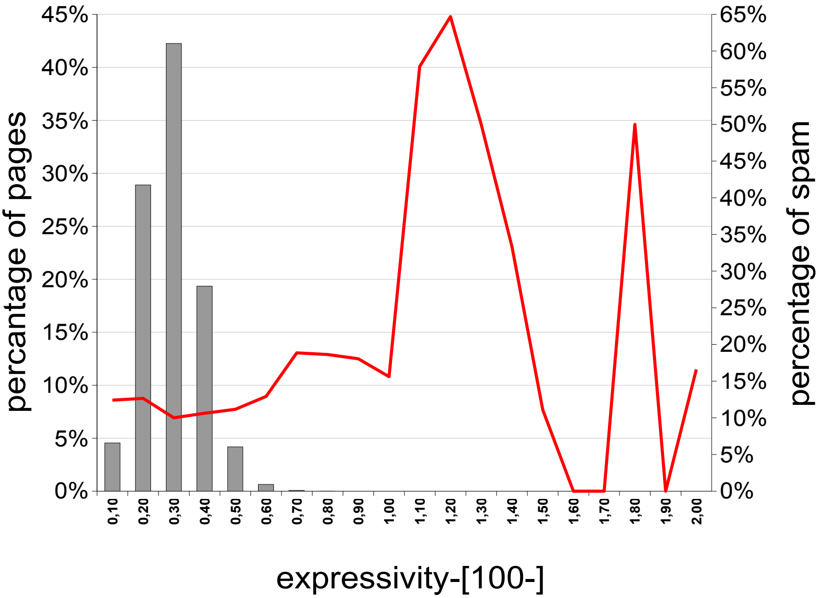

22.5%
20%
17.5%
15%
12.5%
10%
7.5%
5%
2.5%
0%
percantage of pages
18%
17%
16%
15%
14%
13%
12%
11%
10%
9%
8%
7%
6%
5%
4%
3%
percentage of spam
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
lexical validity

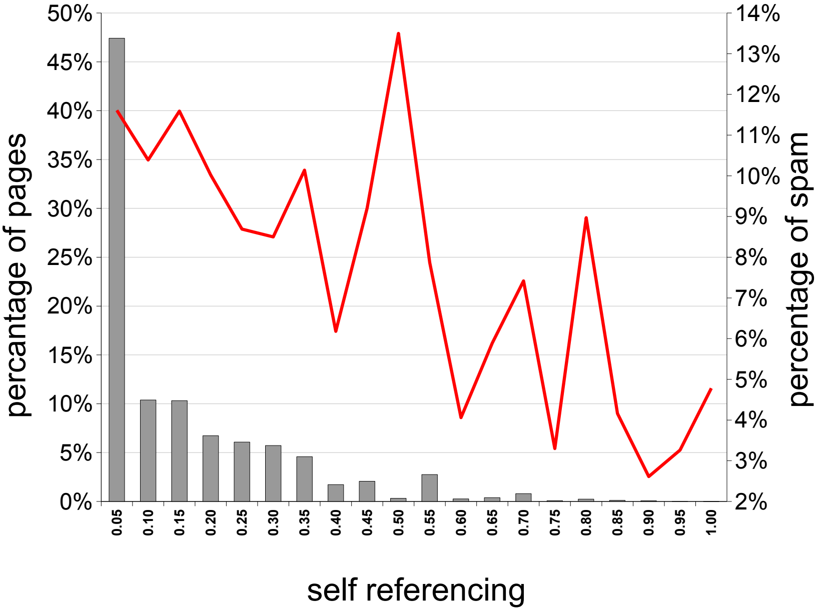

percentage of pages
27.5%
25%
22.5%
20%
17.5%
15%
12.5%
10%
7.5%
5%
2.5%
0%
percentage of spam
19%
18%
17%
16%
15%
14%
13%
12%
11%
10%
9%
8%
7%
6%
5%
4%
3%
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
lexical validity-[100-]


22.5%
20%
17.5%
15%
12.5%
10%
7.5%
5%
2.5%
0%
percentage of pages
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
percentage of spam
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
text-like fraction

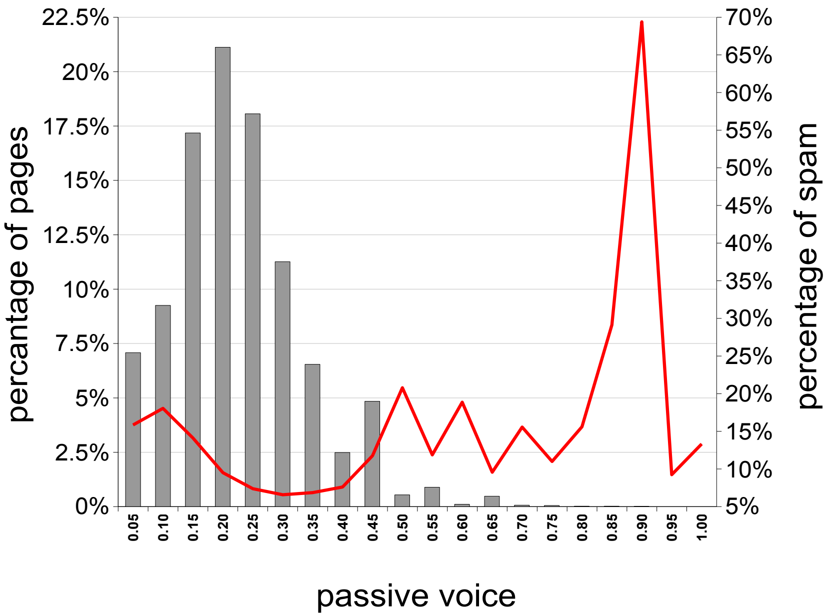

25%
22.5%
20%
17.5%
15%
12.5%
10%
7.5%
5%
2.5%
0%
percentage of pages
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
percentage of spam
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
text-like fraction-[100-]

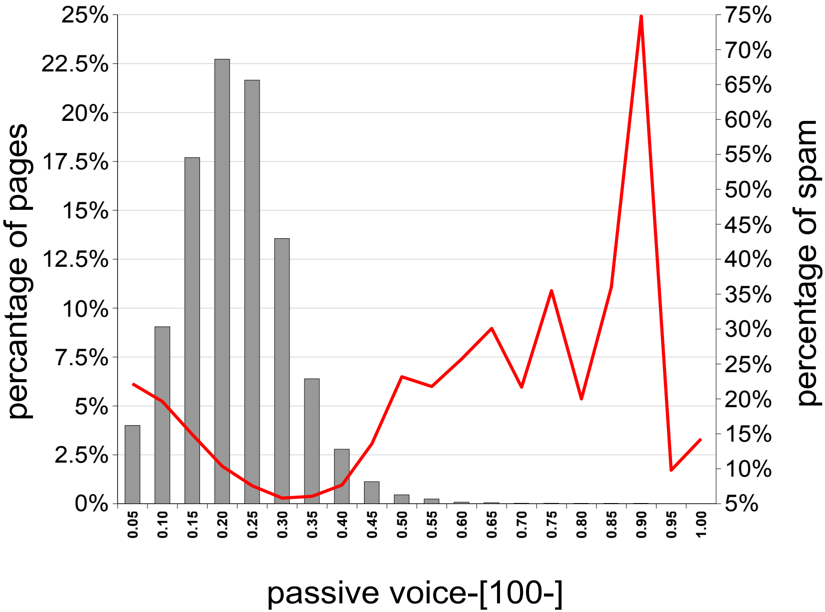

percantage of pages
40%
37,5%
35%
32,5%
30%
27,5%
25%
22,5%
20%
17,5%
15%
12,5%
10%
7,5%
5%
2,5%
0%
percentage of spam
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
0,10 0,20 0,30 0,40 0,50 0,60 0,70 0,80 0,90 1,00 1,10 1,20 1,30 1,40 1,50 1,60 1,70 1,80 1,90 2,00
emotiveness


percantage of pages
percentage of spam
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
0,10
0,20
0,30
0,40
0,50
0,60
0,70
0,80
0,90
1,00
1,10
1,20
1,30
1,40
1,50
1,60
1,70
1,80
1,90
2,00
expressivity-[100-]



percentage of pages
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
percentage of spam
14%
13%
12%
11%
10%
9%
8%
7%
6%
5%
4%
3%
2%
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
self referencing


22.5%
20%
17.5%
15%
12.5%
10%
7.5%
5%
2.5%
0%
percantage of pages
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
percentage of spam
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
passive voice



25%
22.5%
20%
17.5%
15%
12.5%
10%
7.5%
5%
2.5%
0%
percantage of pages
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
percentage of spam
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
passive voice-[100-]

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[Y,S], Luca Becchetti[R], Paolo Boldi[M], Debora Donato[Y], Aristides Gionis[Y], Stefano Leonardi[R], Vanessa Murdock[Y], Massimo Santini[M], Fabrizio Silvestri[P], Sebastiano Vigna[M], Leszek Krupiński[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!

Becchetti, L., Castillo, C., Donato, D., Leonardi, S., and Baeza-Yates, R. (2006).
Using rank propagation and probabilistic counting for link-based spam detection.
In *Proceedings of the Workshop on Web Mining and Web Usage Analysis (WebKDD)*, Pennsylvania, USA. ACM Press.

Chellapilla, K. and Maykov, A. (2007).
A taxonomy of javascript redirection spam.
In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 81–88, New York, NY, USA. ACM Press.

Cohen, W. W. and Kou, Z. (2006).
Stacked graphical learning: approximating learning in markov random fields using very short inhomogeneous markov chains.
Technical report.

Davison, B. D. (2000).
Topical locality in the web.
In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, pages 272–279, Athens, Greece. ACM Press.

Fetterly, D., Manasse, M., and Najork, M. (2004).
Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages.
In *Proceedings of the seventh workshop on the Web and databases (WebDB)*, pages 1–6, Paris, France.

Flajolet, P. and Martin, N. G. (1985).
Probabilistic counting algorithms for data base applications.
Journal of Computer and System Sciences, 31(2):182–209.

Gibson, D., Kumar, R., and Tomkins, A. (2005).
Discovering large dense subgraphs in massive graphs.
In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 721–732. VLDB Endowment.

Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. (2004).
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.

Ntoulas, A., Najork, M., Manasse, M., and Fetterly, D. (2006).
Detecting spam web pages through content analysis.
In *Proceedings of the World Wide Web conference*, pages 83–92, Edinburgh, Scotland.

Palmer, C. R., Gibbons, P. B., and Faloutsos, C. (2002).
ANF: a fast and scalable tool for data mining in massive graphs.
In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 81–90, New York, NY, USA. ACM Press.

Zhou, A., Burgoon, J., Nunamaker, J., and Twitchell, D. (2004).
Automating Linguistics-Based Cues for Detecting Deception of Text-based Asynchronous Computer-Mediated Communication.
Group Decision and Negotiations, 12:81–106.