

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

Luca Becchetti¹, Carlos Castillo¹, Debora Donato¹,
Stefano Leonardi¹ and Ricardo Baeza-Yates²

1. Università di Roma "La Sapienza" – Rome, Italy
2. Yahoo! Research – Barcelona, Spain and Santiago, Chile

August 20th, 2006

- 1 Motivation
- 2 Spam pages characterization
- 3 Truncated PageRank
- 4 Counting supporters
- 5 Experiments
- 6 Conclusions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- 1 Motivation
- 2 Spam pages characterization
- 3 Truncated PageRank
- 4 Counting supporters
- 5 Experiments
- 6 Conclusions

What is on the Web?

Using rank
propagation and
Probabilistic
counting for
Link-Based
Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages
characterization

Truncated
PageRank

Counting
supporters

Experiments

Conclusions

Information

What is on the Web?

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Information + Porn

Web spam (keywords + links)

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

side effects, at strength of erection viagra levitra cialis, discount viagra buy viagra buy viagra viagradrugs.net, to cialis lawsuit, dirt cheap viagra, in sex discount cialis generic cialis bluepilled.com, herbal alternative viagra, for cialis marijuana, sublingual viagra.

Viagra users, will viagra facts cialis line prescription, buy viagra online viagra side effects natural alternative viagra, has cialis generic viagra generic cialis cialis cum-with-us.com, viagra discount, this brand name cialis, herbal viagra alternative free viagra buying deal viagradrugs.net cheapest price viagra cheap viagra uk free viagra viagra online pills pills viagradrugs.net, silagra weight loss generic viagra cialis cum-with-us.com, viagra blindness viagra prescription.

Amsterdam viagra sexshops viagra prescription for woman viagra online pharmacy, is cialis ordering online, viagra suppliers cocaine and viagra sex experiences viagra generico impotencia, cialis official website, viagra cheap generic cheap viagra natural viagra, will ciali, whats the chemical name for the drug viagra, are cialis and grapefruit, homemade viagra, has herbal cialis, strength of erection viagra levitra cialis.

Viagra for women, has viagra cost lowest prices viagra, at cialis eli lilly, non prescription viagra, am cialis on line, viagra for women viagra expiration cialis fda approval, compare viagra and levitra viagra discount viagra cialis levitra, viagra online cheap cialis no prescription, 180 mg viagra levitra vs viagra uk viagra viagra sample, am generic cialis minuteviagra cum-with-us.com, free viagra online.

Herbal viagra samples, to order viagra visit your doctor online viagra substitute side effects from viagra cheapest price viagra, by cialis soft tab, mail order viagra, for cialis store, british viagra, is cialis fedex overnight, viagra suppliers cialis herbalsubstitute.com, whats the chemical name for the drug viagra herbal viagra viagra info

- ♦ [generic viagra](#)
- ♦ [buy viagra](#)
- ♦ [viagra alternative](#)
- ♦ [herbal viagra](#)
- ♦ [cheap viagra](#)
- ♦ [viagra online](#)
- ♦ [buy viagra online](#)
- ♦ [order viagra](#)
- ♦ [order viagra online](#)
- ♦ [Viagra](#)
- ♦ [natural viagra](#)
- ♦ [viagra pill](#)
- ♦ [free viagra samples](#)
- ♦ [discount viagra](#)
- ♦ [female viagra](#)
- ♦ [viagra](#)

Web spam (mostly keywords)

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

[smart movie converter 2.72 registration key as nokia6600](#)

crack diablo2, kontakt 2 .exe, download crack norton 2006 online, downloand snagit, telecharger canopu nothing else mather/ lyric, silent hill 3 no-dvd crack, Protocol v7 Update-Vengeance.rar download, kn windows validation crack download

ftp downloads spanish,

total commander hack,

plus superpack key,

crack de need ford

speed underground,

manual diablo 2,

Rapture_dhol_mix.mp3,

advance 3gp converter,

pacific assault torrent,

Fruity Loops 4.5.1

Demo, russian mohaa

skin, telechargement de

crack battlefield 2

key generator for easy cd-da extractor v9, fine rider

8.0, download demo fifa street pc, soundtrack

moulin rouge free mp3, crazy froog popcorn, Utah

Saints - Take On The Theme From Mortal Kombat

mp3, cunter-strike password, download free game

sex, SYSTEM OF DOWN DIRECTORY

PARENT MPEG

wifi download key generator wap, speedstream

feature activation, command conquer generals key

gens, nero vision directx9.0 download, swift 3D trial

cracks, winamp crack licens, mobil msn sis, nero

burning room 6.0.0.19, Deep Silver Keygen, the

sims makin ` magik exploration pack Flash Get,

DesktopX) 3.1 (keygen l serial, application

architecture control, download ogc quake3

free gemu,

telecharger

cdkeys, care day

home question,

Remote S60

software

cracked, nt print

server,

SWF2V video

Plugin for

Adobe Premiere

Pro cracked, ps2

secret code

download

webcam

lv-c300,

[download counter strike 1.5](#)

Search engine?

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions



Top Searches:

- » Acne
- » Weight Loss Pills
- » Debt Consolidation
- » Loan
- » Domain Names
- » Advertising
- » Online Pharmacy
- » Home Loan
- » Dedicated Server
- » Car Rental
- » Adipex
- » Levitra
- » Online Poker
- » Work At Home
- » Propecia
- » Consolidate Debt
- » Mortgage Rates
- » Online Craps
- » Vegas Casinos
- » Buy Ionamin

lava soft

php script

top soft

java script

MP3

Top Web Results

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 sportsbook in North America
<http://www.sportsinteraction.com>
2. **AnteUp GamblingLinks.com - Safe Online Casinos**
Links to safe and secure online casino gambling and sports betting including reviews, ne
<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 f
rules and strategy. Links to the Best Casinos
<http://www.fastfreecash.net>
4. **AnteUp GamblingLinks.com - Safe Online Casinos**

Fake search engine

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization


Truncated PageRank

Counting supporters

Experiments

Conclusions

→ Bookmark → Home Page → Home


SOFT SEARCH

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

lava soft php script top soft java script MP3

Top Web Results

Results 1-16 containing "1293kasd132ka0sd1kj239asd123"

- 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/>
- 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/>
- 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>
- Renting a Birthday Party Limousine is Sexy**
What better way to surprise your loved one on their special day than with a birthday party?
<http://partybusrental.info>

Problem: “normal” pages that are spam

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

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.](#)

Problem: "normal" pages that are spam

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

The screenshot shows a Mozilla Firefox browser window with the address bar displaying `http://www.home-security-webpage.com/home-security-system-separate-blasts-kill-ne`. The page content includes:

- Header: Home Security Webpage
- Navigation: File, Edit, View, Go, Bookmarks, Tools, Help
- Search: My Yahoo!, SK posts, com, Ecosofia, com
- Web Spam Test Collections: Home Security Webpage..., (Untitled)
- Advertisements:
 - Ads by Goooooogle: Advertise on this site
 - 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.stopstechld.com
- Archived Entry:
 - Post Date: Tuesday, Nov 22nd, 2005 at 2:03 pm
 - Category: Uncategorized
 - Do More: You can trackback from your own site.
- Additional Ads:
 - Ads by Goooooogle: Prevent Home Burglary: Home burglary is rampant. Read all about security systems. www.for-the-touchdow
 - Security Industry News: Latest on CCTV, loss prevention, access control & more for pros
- 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 Explosion Video from a security camera at the Hamra Hotel in Baghdad look at the fallen troops' home towns, ages, service categories and other
- Rood girl's game of strip

The browser status bar at the bottom shows "Find: filters" and "Done".

Some content is introduced

Problem: borderline pages

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Medical Term brought to you by Medkuz - Mozilla Firefox

File Edit View Go Bookmarks Tools Help None

My Yahoo! SK posts com Ecosofia com Digo

http://medkuz.com/Medical_Term.php?id=69&term=Hepatitis

Web Spa... Home Sec... (Untitled) adsense... AdsBlack... Medical ... Hepatitis

MedKuz, Health and Medical Info

Hepatitis

Hepatitis is a gastroenterological disease, featuring inflammation of the liver. The clinical signs and prognosis, as well as the therapy, depend on the cause. Hepatitis is characterised by abdominal pain, fever, hepatomegaly (enlarged liver) and jaundice (icterus). Some chronic forms of hepatitis show very few of these signs and only present when the longstanding inflammation has led to the replacement of liver cells by connective tissue; the result is cirrhosis. Certain liver function tests can also indicate hepatitis

Ads by Google

- [Lactose intolerance diet](#)
- [Cure Diarrhea](#)
- [Lactaid](#)
- [Diarrhea Remedy](#)

Ads by Google

- [Lactose intolerance](#)
- [Diarrhea & Bowel Problems](#)

Find: filters Find Next Find Previous Highlight all Match case

Done Disabled Proxy: N

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- 1 Motivation
- 2 Spam pages characterization**
- 3 Truncated PageRank
- 4 Counting supporters
- 5 Experiments
- 6 Conclusions

Definitions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Any deliberate action that is meant to trigger an unjustifiably favorable relevance or importance for some Web page, considering the page's true value
[Gyöngyi and Garcia-Molina, 2005]

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Any deliberate action that is meant to trigger
an unjustifiably favorable relevance or importance
for some Web page, considering the page's true value
[Gyöngyi and Garcia-Molina, 2005]

any attempt to deceive a search engine's relevancy algorithm

or simply

anything that would not be done
if search engines did not exist.
[Perkins, 2001]

Link farms

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

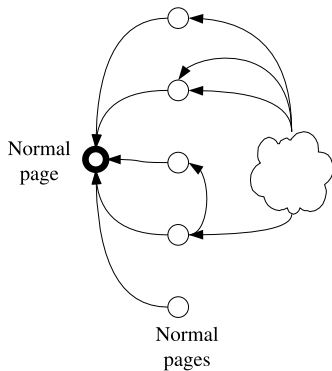
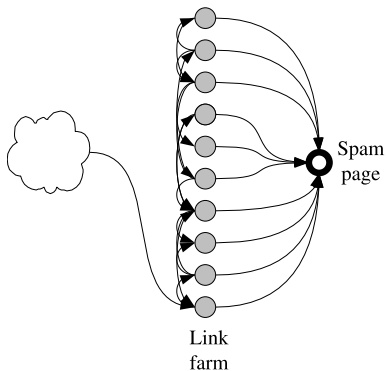
Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions



Link farms

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

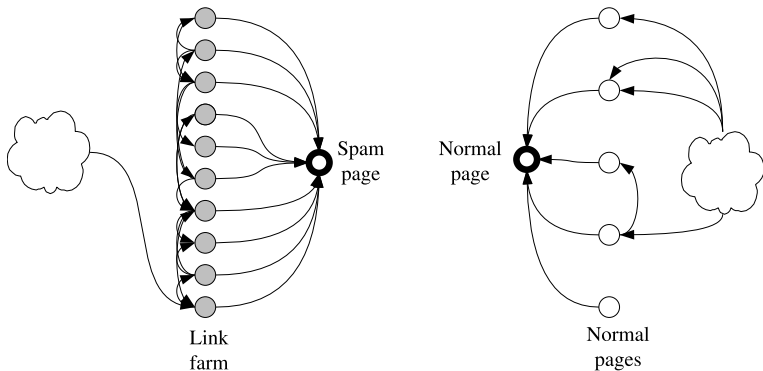
Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions



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

Link-based spam

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

[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”

Link-based spam

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

[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”

Research goal

Statistical analysis of link-based spam

Idea: count “supporters” at different distances

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

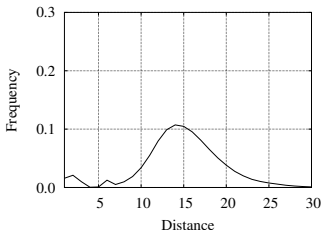
Counting supporters

Experiments

Conclusions

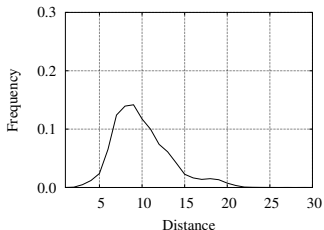
Number of different nodes at a given distance:

.UK 18 mill. nodes



Average distance
14.9 clicks

.EU.INT 860,000 nodes



Average distance
10.0 clicks

High and low-ranked pages are different

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

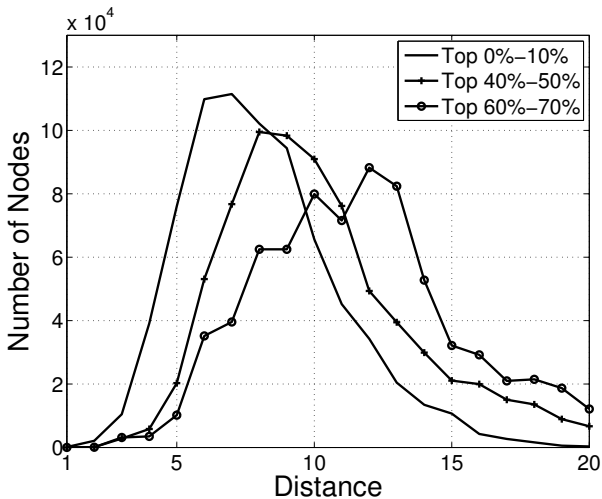
Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions



High and low-ranked pages are different

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

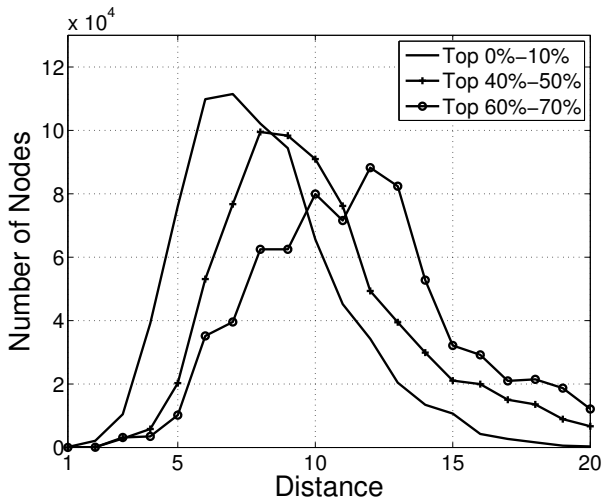
Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions



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

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Graph algorithms

All shortest paths, centrality, betweenness, clustering coefficient...

Streamed algorithms

Breadth-first and depth-first search

Count of neighbors

Symmetric algorithms

(Strongly) connected components

Approximate count of neighbors

PageRank, Truncated PageRank, Linear Rank

HITS, Salsa, TrustRank

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- 1 Motivation
- 2 Spam pages characterization
- 3 Truncated PageRank**
- 4 Counting supporters
- 5 Experiments
- 6 Conclusions

General functional ranking

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Let \mathbf{P} the row-normalized version of the citation matrix of a graph $G = (V, E)$

A **functional ranking** [Baeza-Yates et al., 2006] is a link-based ranking algorithm to compute a scoring vector \mathbf{W} of the form:

$$\mathbf{W} = \sum_{t=0}^{\infty} \frac{\text{damping}(t)}{N} \mathbf{P}^t .$$

General functional ranking

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Let \mathbf{P} the row-normalized version of the citation matrix of a graph $G = (V, E)$

A **functional ranking** [Baeza-Yates et al., 2006] is a link-based ranking algorithm to compute a scoring vector \mathbf{W} of the form:

$$\mathbf{W} = \sum_{t=0}^{\infty} \frac{\text{damping}(t)}{N} \mathbf{P}^t .$$

There are many choices for $\text{damping}(t)$, including simply a linear function that is as good as PageRank in practice

General functional ranking

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Let \mathbf{P} the row-normalized version of the citation matrix of a graph $G = (V, E)$

A **functional ranking** [Baeza-Yates et al., 2006] is a link-based ranking algorithm to compute a scoring vector \mathbf{W} of the form:

$$\mathbf{W} = \sum_{t=0}^{\infty} \frac{\text{damping}(t)}{N} \mathbf{P}^t .$$

There are many choices for $\text{damping}(t)$, including simply a linear function that is as good as PageRank in practice

$$\text{damping}(t) = (1 - \alpha)\alpha^t$$

Truncated PageRank

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

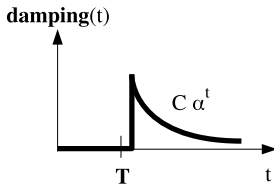
Truncated PageRank

Counting supporters

Experiments

Conclusions

Reduce the direct contribution of the first levels of links:



$$damping(t) = \begin{cases} 0 & t \leq T \\ C\alpha^t & t > T \end{cases}$$

Truncated PageRank

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

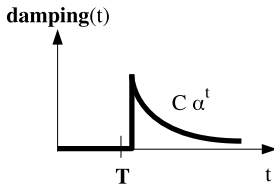
Truncated PageRank

Counting supporters

Experiments

Conclusions

Reduce the direct contribution of the first levels of links:



$$damping(t) = \begin{cases} 0 & t \leq T \\ C\alpha^t & t > T \end{cases}$$

✓ No extra reading of the graph after PageRank

General algorithm

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Require: N : number of nodes, $0 < \alpha < 1$: damping factor, $T \geq -1$: distance for truncation

```
1: for  $i : 1 \dots N$  do {Initialization}
2:    $R[i] \leftarrow (1 - \alpha) / ((\alpha^{T+1})N)$ 
3:   if  $T \geq 0$  then
4:      $Score[i] \leftarrow 0$ 
5:   else {Calculate normal PageRank}
6:      $Score[i] \leftarrow R[i]$ 
7:   end if
8: end for
```


General algorithm

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

```
Require: N: number of nodes,  $0 < \alpha < 1$ : damping factor,  $T \geq -1$ : distance for truncation
1: for  $i : 1 \dots N$  do {Initialization}
2:    $R[i] \leftarrow (1 - \alpha) / ((\alpha^{T+1})N)$ 
3:   if  $T \geq 0$  then
4:      $Score[i] \leftarrow 0$ 
5:   else {Calculate normal PageRank}
6:      $Score[i] \leftarrow R[i]$ 
7:   end if
8: end for
9: distance = 1
10: while not converged do
11:   Aux  $\leftarrow 0$ 
12:   for src : 1 ... N do {Follow links in the graph}
13:     for all link from src to dest do
14:       Aux[dest]  $\leftarrow$  Aux[dest] + R[src]/outdegree(src)
15:     end for
16:   end for
17:   for  $i : 1 \dots N$  do {Apply damping factor  $\alpha$ }
18:      $R[i] \leftarrow Aux[i] \times \alpha$ 
19:     if distance > T then {Add to ranking value}
20:        $Score[i] \leftarrow Score[i] + R[i]$ 
21:     end if
22:   end for
23:   distance = distance + 1
24: end while
```

Truncated PageRank vs PageRank

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

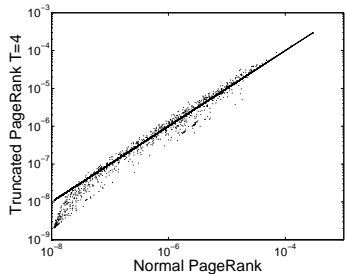
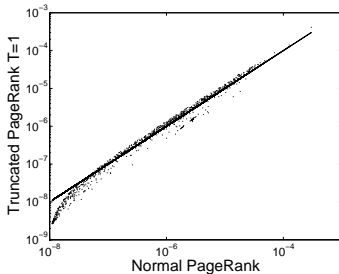
Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions



Comparing PageRank and Truncated PageRank with $T = 1$ and $T = 4$.

The correlation is high and decreases as more levels are truncated.

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- 1 Motivation
- 2 Spam pages characterization
- 3 Truncated PageRank
- 4 Counting supporters**
- 5 Experiments
- 6 Conclusions

Probabilistic counting

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

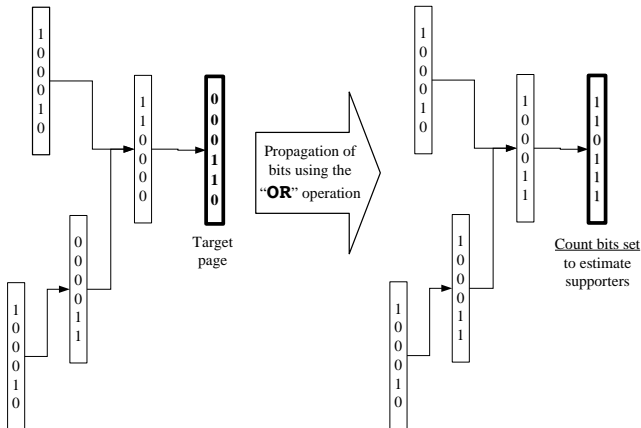
Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions



Probabilistic counting

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

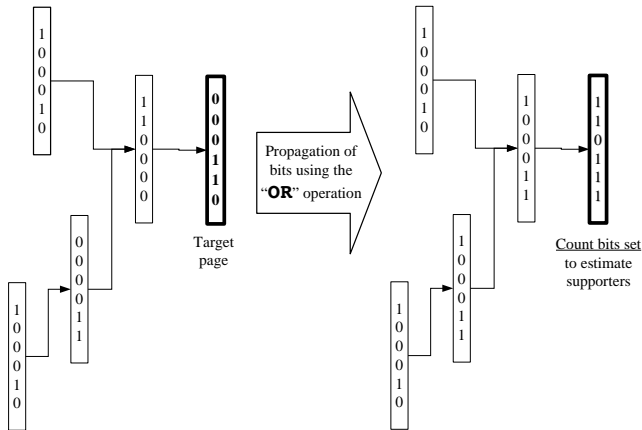
Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions



Improvement of ANF algorithm [Palmer et al., 2002] based on probabilistic counting [Flajolet and Martin, 1985]

General algorithm

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Require: N: number of nodes, d: distance, k: bits

- 1: **for** node : 1 ... N, bit: 1 ... k **do**
- 2: INIT(node,bit)
- 3: **end for**

General algorithm

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Require: N: number of nodes, d: distance, k: bits

```
1: for node : 1 ... N, bit: 1 ... k do
2:   INIT(node,bit)
3: end for
4: for distance : 1 ... d do {Iteration step}
5:   Aux  $\leftarrow \mathbf{0}_k$ 
6:   for src : 1 ... N do {Follow links in the graph}
7:     for all links from src to dest do
8:       Aux[dest]  $\leftarrow$  Aux[dest] OR V[src,·]
9:     end for
10:  end for
11:  V  $\leftarrow$  Aux
12: end for
```

General algorithm

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

```
Require: N: number of nodes, d: distance, k: bits
1: for node : 1 ... N, bit: 1 ... k do
2:   INIT(node,bit)
3: end for
4: for distance : 1 ... d do {Iteration step}
5:   Aux  $\leftarrow \mathbf{0}_k$ 
6:   for src : 1 ... N do {Follow links in the graph}
7:     for all links from src to dest do
8:       Aux[dest]  $\leftarrow$  Aux[dest] OR V[src,·]
9:     end for
10:  end for
11:  V  $\leftarrow$  Aux
12: end for
13: for node: 1 ... N do {Estimate supporters}
14:   Supporters[node]  $\leftarrow$  ESTIMATE( V[node,·] )
15: end for
16: return Supporters
```


Our estimator

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Initialize all bits to one with probability ϵ

Our estimator

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Initialize all bits to one with probability ϵ
by the independence of the i - th component X_i 's we have,

$$\mathbf{P}[X_i = 1] = 1 - (1 - \epsilon)^{\text{neighbors}(\text{node})},$$

Our estimator

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Initialize all bits to one with probability ϵ
by the independence of the i -th component X_i 's we have,

$$\mathbf{P}[X_i = 1] = 1 - (1 - \epsilon)^{\text{neighbors}(\text{node})},$$

$$\text{Estimator: } \text{neighbors}(\text{node}) = \log_{(1-\epsilon)} \left(1 - \frac{\text{ones}(\text{node})}{k} \right)$$

Our estimator

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Initialize all bits to one with probability ϵ
by the independence of the i - *th* component X_i 's we have,

$$\mathbf{P}[X_i = 1] = 1 - (1 - \epsilon)^{\text{neighbors}(\text{node})},$$

Estimator: $\text{neighbors}(\text{node}) = \log_{(1-\epsilon)} \left(1 - \frac{\text{ones}(\text{node})}{k} \right)$

Problem: $\text{neighbors}(\text{node})$ can vary by orders of magnitudes as node varies.

Our estimator

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Initialize all bits to one with probability ϵ
by the independence of the i -th component X_i 's we have,

$$\mathbf{P}[X_i = 1] = 1 - (1 - \epsilon)^{\text{neighbors}(\text{node})},$$

Estimator: $\text{neighbors}(\text{node}) = \log_{(1-\epsilon)} \left(1 - \frac{\text{ones}(\text{node})}{k} \right)$

Problem: $\text{neighbors}(\text{node})$ can vary by orders of magnitudes as node varies.

This means that for some values of ϵ , the computed value of $\text{ones}(\text{node})$ might be k (or 0, depending on $\text{neighbors}(\text{node})$) with relatively high probability.

Adaptive estimator

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

if we knew $\text{neighbors}(\text{node})$ and chose $\epsilon = \frac{1}{\text{neighbors}(\text{node})}$ we would get:

$$\text{ones}(\text{node}) \simeq \left(1 - \frac{1}{e}\right) k \simeq 0.63k,$$

Adaptive estimation

Repeat the above process for $\epsilon = 1/2, 1/4, 1/8, \dots$, and look for the transitions from more than $(1 - 1/e)k$ ones to less than $(1 - 1/e)k$ ones.

Convergence

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

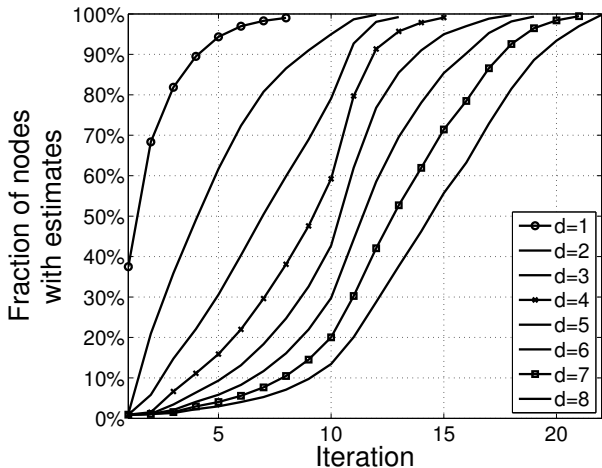
Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions



15 iterations for estimating the neighbors at distance 4 or less

Convergence

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

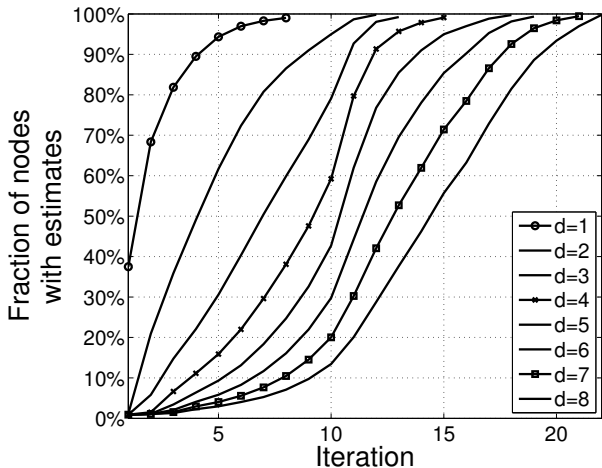
Spam pages
characterization

Truncated
PageRank

Counting
supporters

Experiments

Conclusions



15 iterations for estimating the neighbors at distance 4 or less
less than 25 iterations for all distances up to 8.

Error rate

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

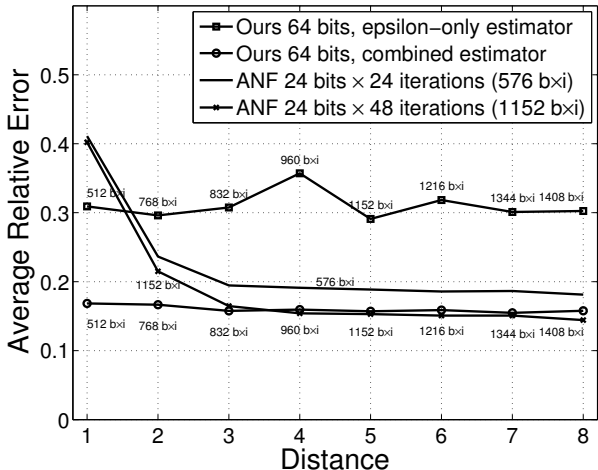
Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions



Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- 1 Motivation
- 2 Spam pages characterization
- 3 Truncated PageRank
- 4 Counting supporters
- 5 Experiments**
- 6 Conclusions

Test collection

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

U.K. collection

18.5 million pages downloaded from the .UK domain

5,344 hosts manually classified (6% of the hosts)

Test collection

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

U.K. collection

18.5 million pages downloaded from the .UK domain

5,344 hosts manually classified (6% of the hosts)

Classified entire hosts:

- ✓ **A few hosts are mixed:** spam and non-spam pages
- ✗ **More coverage:** sample covers 32% of the pages

Automatic classifier

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages
characterization

Truncated
PageRank

Counting
supporters

Experiments

Conclusions

We extracted (for the home page and the page with maximum PageRank) PageRank, Truncated PageRank at 2...4, Supporters at 2...4

Automatic classifier

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

We extracted (for the home page and the page with maximum PageRank) PageRank, Truncated PageRank at 2...4, Supporters at 2...4

We measured:

$$\text{Precision} = \frac{\# \text{ of spam hosts classified as spam}}{\# \text{ of hosts classified as spam}}$$

$$\text{Recall} = \frac{\# \text{ of spam hosts classified as spam}}{\# \text{ of spam hosts}} .$$

Automatic classifier

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

We extracted (for the home page and the page with maximum PageRank) PageRank, Truncated PageRank at 2...4, Supporters at 2...4

We measured:

$$\text{Precision} = \frac{\# \text{ of spam hosts classified as spam}}{\# \text{ of hosts classified as spam}}$$

$$\text{Recall} = \frac{\# \text{ of spam hosts classified as spam}}{\# \text{ of spam hosts}} .$$

and the two types of errors in spam classification

$$\text{False positive rate} = \frac{\# \text{ of normal hosts classified as spam}}{\# \text{ of normal hosts}}$$

$$\text{False negative rate} = \frac{\# \text{ of spam hosts classified as normal}}{\# \text{ of spam hosts}} .$$

Single-technique classifier

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Classifier based on TrustRank: uses as features the PageRank, the estimated non-spam mass, and the estimated non-spam mass divided by PageRank.

Classifier based on Truncated PageRank: uses as features the PageRank, the Truncated PageRank with truncation distance $t = 2, 3, 4$ (with $t = 1$ it would be just based on in-degree), and the Truncated PageRank divided by PageRank.

Classifier based on Estimation of Supporters: uses as features the PageRank, the estimation of supporters at a given distance $d = 2, 3, 4$, and the estimation of supporters divided by PageRank.

Comparison of single-technique classifier ($M=5$)

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Classifiers (pruning with $M = 5$)	Spam class		False	False
	Prec.	Recall	Pos.	Neg.
TrustRank	0.82	0.50	2.1%	50%
Trunc. PageRank $t = 2$	0.85	0.50	1.6%	50%
Trunc. PageRank $t = 3$	0.84	0.47	1.6%	53%
Trunc. PageRank $t = 4$	0.79	0.45	2.2%	55%
Est. Supporters $d = 2$	0.78	0.60	3.2%	40%
Est. Supporters $d = 3$	0.83	0.64	2.4%	36%
Est. Supporters $d = 4$	0.86	0.64	2.0%	36%

Comparison of single-technique classifier ($M=30$)

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Classifiers (pruning with $M = 30$)	Spam class		False	False
	Prec.	Recall	Pos.	Neg.
TrustRank	0.80	0.49	2.3%	51%
Trunc. PageRank $t = 2$	0.82	0.43	1.8%	57%
Trunc. PageRank $t = 3$	0.81	0.42	1.8%	58%
Trunc. PageRank $t = 4$	0.77	0.43	2.4%	57%
Est. Supporters $d = 2$	0.76	0.52	3.1%	48%
Est. Supporters $d = 3$	0.83	0.57	2.1%	43%
Est. Supporters $d = 4$	0.80	0.57	2.6%	43%

Combined classifier

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Pruning	Rules	Spam class		False	False
		Precision	Recall	Pos.	Neg.
M=5	49	0.87	0.80	2.0%	20%
M=30	31	0.88	0.76	1.8%	24%
No pruning	189	0.85	0.79	2.6%	21%

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- 1 Motivation
- 2 Spam pages characterization
- 3 Truncated PageRank
- 4 Counting supporters
- 5 Experiments
- 6 Conclusions**

Summary of classifiers

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

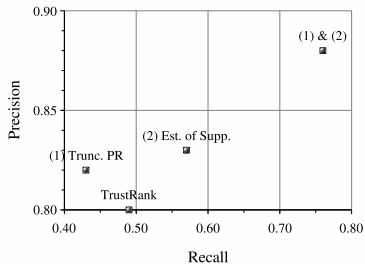
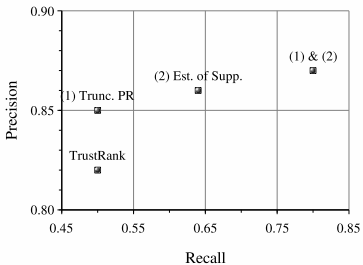
Truncated PageRank

Counting supporters

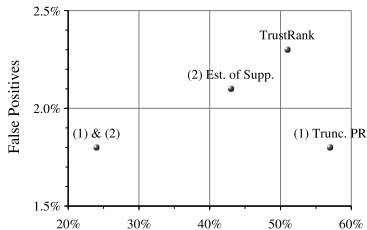
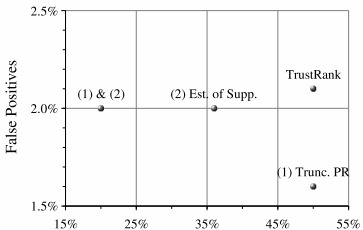
Experiments

Conclusions

(a) Precision and recall of spam detection



(b) Error rates of the spam classifiers



Conclusions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

✔ Link-based statistics to detect 80% of spam

Conclusions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- ✓ Link-based statistics to detect 80% of spam
- ✗ No magic bullet in link analysis

Conclusions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- ✓ Link-based statistics to detect 80% of spam
- ✗ No magic bullet in link analysis
- ✗ Precision still low compared to e-mail spam filters

Conclusions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- ✓ Link-based statistics to detect 80% of spam
- ✗ No magic bullet in link analysis
- ✗ Precision still low compared to e-mail spam filters
- ✓ Measure both home page and max. PageRank page

Conclusions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- ✓ Link-based statistics to detect 80% of spam
- ✗ No magic bullet in link analysis
- ✗ Precision still low compared to e-mail spam filters
- ✓ Measure both home page and max. PageRank page
- ✓ Host-based counts are important

Conclusions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- ✓ Link-based statistics to detect 80% of spam
- ✗ No magic bullet in link analysis
- ✗ Precision still low compared to e-mail spam filters
- ✓ Measure both home page and max. PageRank page
- ✓ Host-based counts are important

Conclusions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

- ✓ Link-based statistics to detect 80% of spam
- ✗ No magic bullet in link analysis
- ✗ Precision still low compared to e-mail spam filters
- ✓ Measure both home page and max. PageRank page
- ✓ Host-based counts are important

Next step: combine link analysis and content analysis

Using rank
propagation and
Probabilistic
counting for
Link-Based
Spam Detection

L. Becchetti,
C. Castillo,
D. Donato,
S. Leonardi and
R. Baeza-Yates

Motivation

Spam pages
characterization

Truncated
PageRank

Counting
supporters

Experiments

Conclusions

Thank you!



Baeza-Yates, R., Boldi, P., and Castillo, C. (2006).
Generalizing PageRank: Damping functions for link-based ranking algorithms.

In Proceedings of SIGIR, Seattle, Washington, USA. ACM Press.



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.



Benczúr, A. A., Csalogány, K., Sarlós, T., and Uher, M. (2005).

Spamrank: fully automatic link spam detection.

In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, Chiba, Japan.



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. and Garcia-Molina, H. (2005).

Web spam taxonomy.

In First International Workshop on Adversarial Information Retrieval on the Web.



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

Combating web spam with trustrank.

In Proceedings of the Thirtieth International Conference on Very Large Data Bases (VLDB), pages 576–587, Toronto, Canada. Morgan Kaufmann.



Newman, M. E., Strogatz, S. H., and Watts, D. J. (2001).

Random graphs with arbitrary degree distributions and their applications.

Phys Rev E Stat Nonlin Soft Matter Phys, 64(2 Pt 2).



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.



Perkins, A. (2001).

The classification of search engine spam.

Available online at

<http://www.silverdisc.co.uk/articles/spam-classification/>.