Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

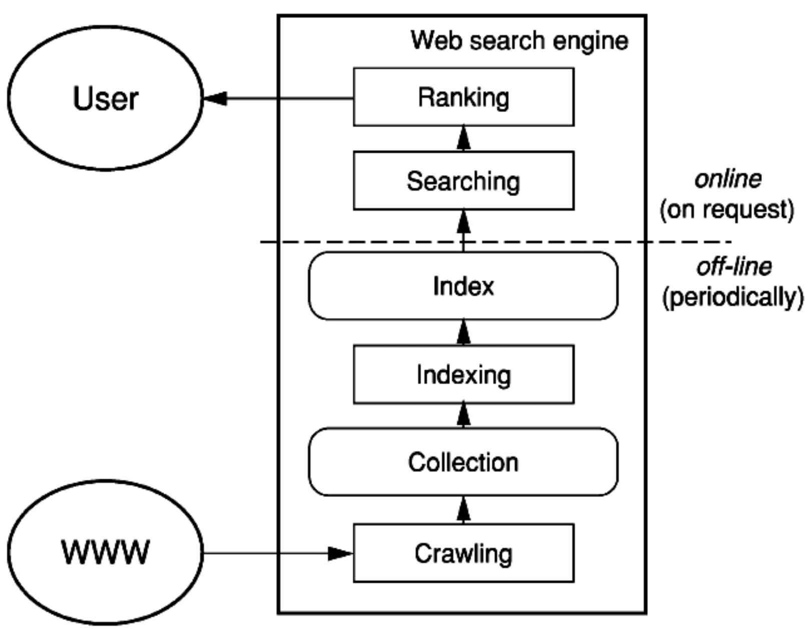
Web Spam Detection

Ricardo Baeza-Yates[1,3]
ricardo@baeza.cl
With: L. Becchetti[2], P. Boldi[5], C. Castillo[1], D. Donato[1], A. Gionis[1], S. Leonardi[2], V.Murdock[1], M. Santini[5], F. Silvestri[4], S. Vigna[5]

1. Yahoo! Research Barcelona – Catalunya, Spain
2. Università di Roma "La Sapienza" – Rome, Italy
3. Yahoo! Research Santiago – Chile
4. ISTI-CNR –Pisa,Italy
5. Università degli Studi di Milano – Milan, Italy

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Previous: how search engines work



Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Search engine: issues

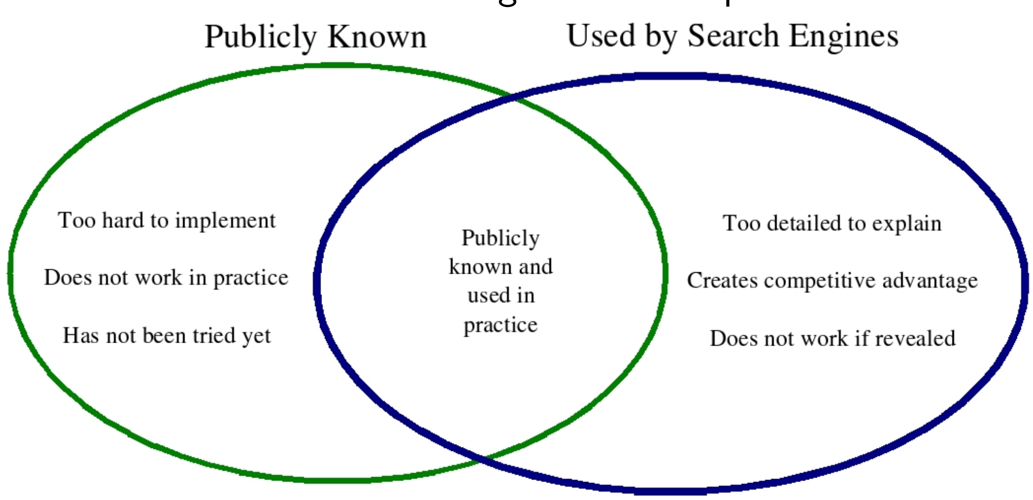
- Scalability (crawling, indexing, searching, ranking)
- Relevance (query to document match)
- Static ranking (content quality)
- Incentives for cheating ($)

Web Spam Detection
R. Baeza-Yates

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

This is a talk about academic research!

Tools for dealing with Web Spam



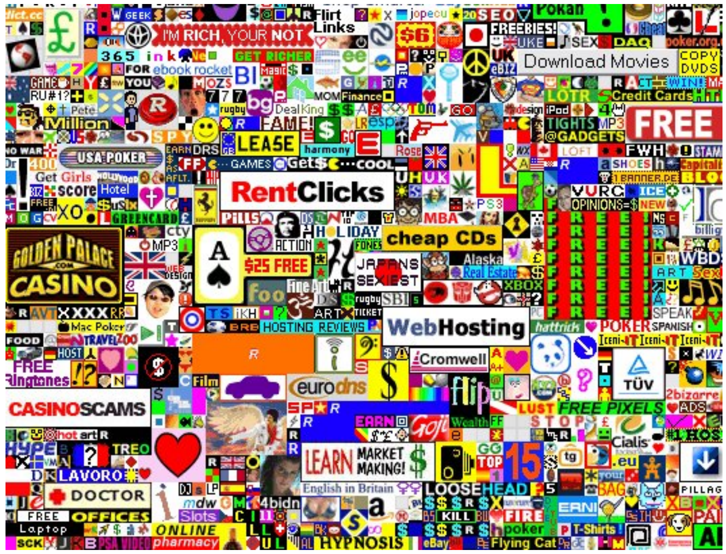
1. Web Spam
2. Web Spam Detection
3. A Reference Collection
4. Web Links
5. Topological Web Spam
6. Counting of Supporters
7. Content-based Spam detection
8. Web Topology
9. Conclusions

Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

1. Web Spam
2. Web Spam Detection
3. A Reference Collection
4. Web Links
5. Topological Web Spam
6. Counting of Supporters
7. Content-based Spam detection
8. Web Topology
9. Conclusions

The Web

"The sum of all human knowledge plus porn" – Robert Gilbert



Graphic: www.milliondollarhomepage.com

Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Adversarial IR Issues on the Web

- Link spam
- Content spam
- Cloaking
- Comment/forum/wiki spam
- Spam-oriented blogging
- Click fraud $\times 2$
- Reverse engineering of ranking algorithms
- Web content filtering
- Advertisement blocking
- Stealth crawling
- Malicious tagging
- ... more?

Opportunities for Web spam

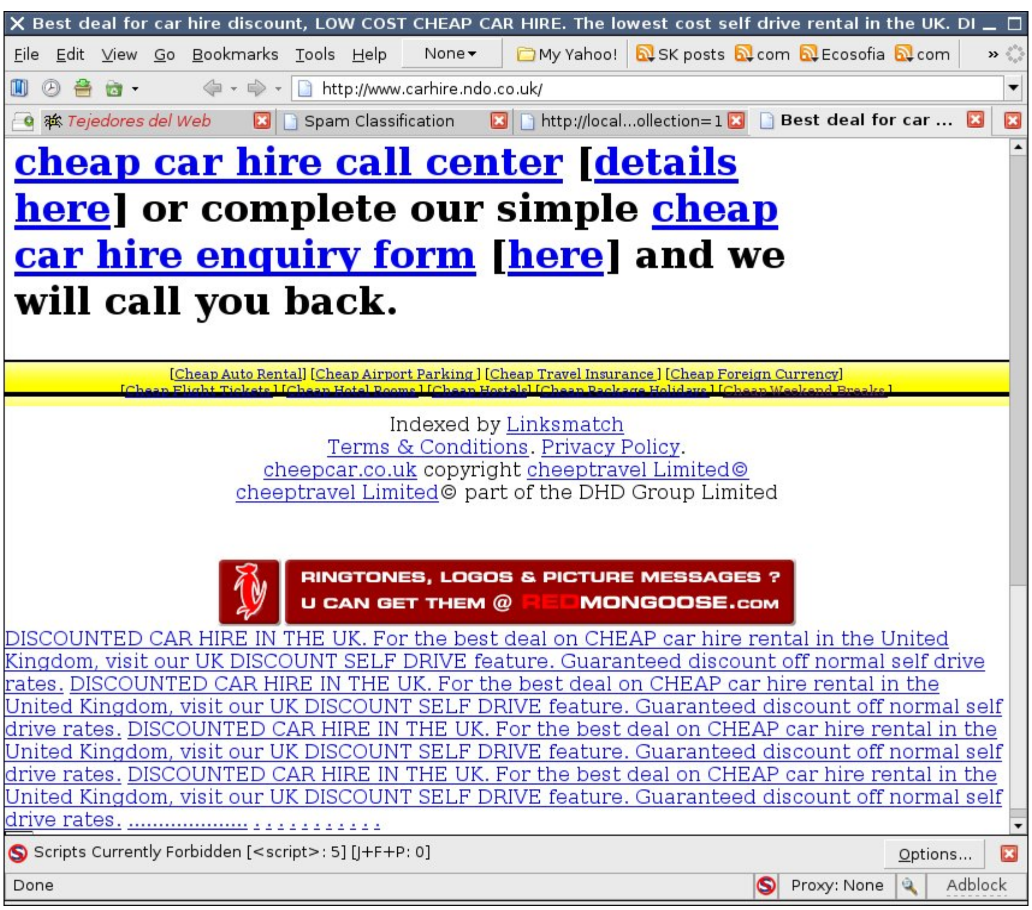
- Spamdexing
  - Keyword stuffing
  - Link farms
  - Spam blogs (splogs)
  - Cloaking

Adversarial relationship

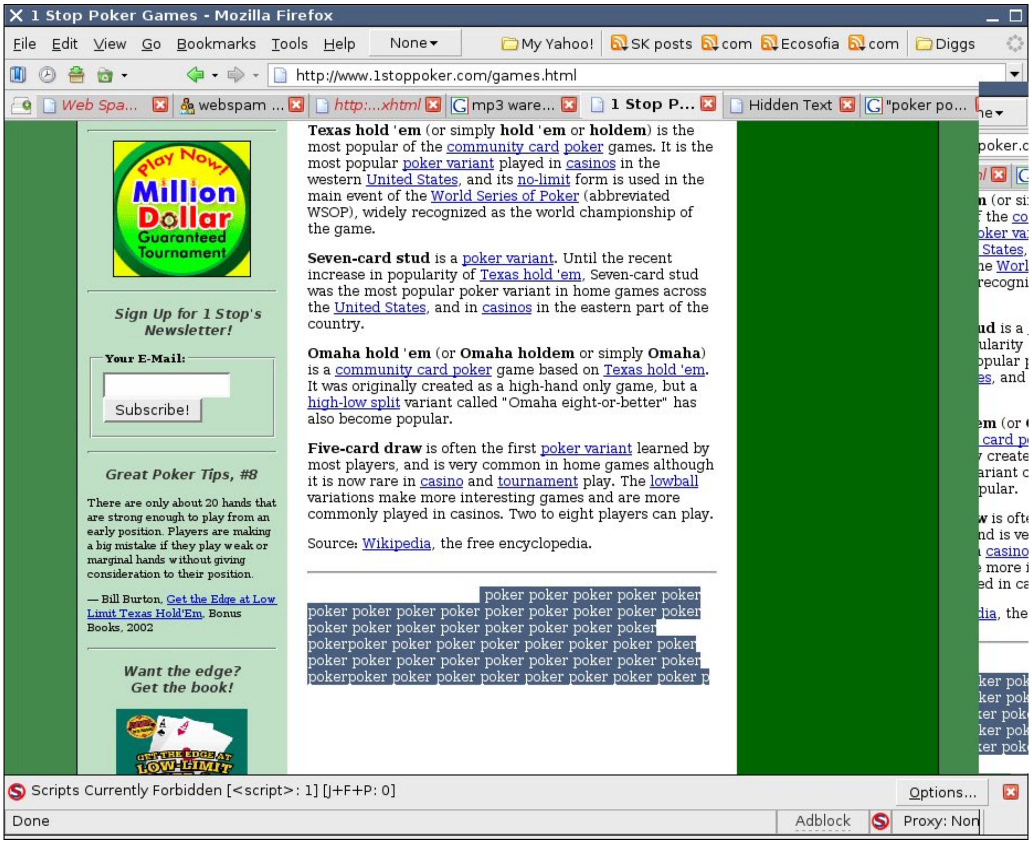
Every undeserved gain in ranking for a spammer, is a loss of precision for the search engine.

Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

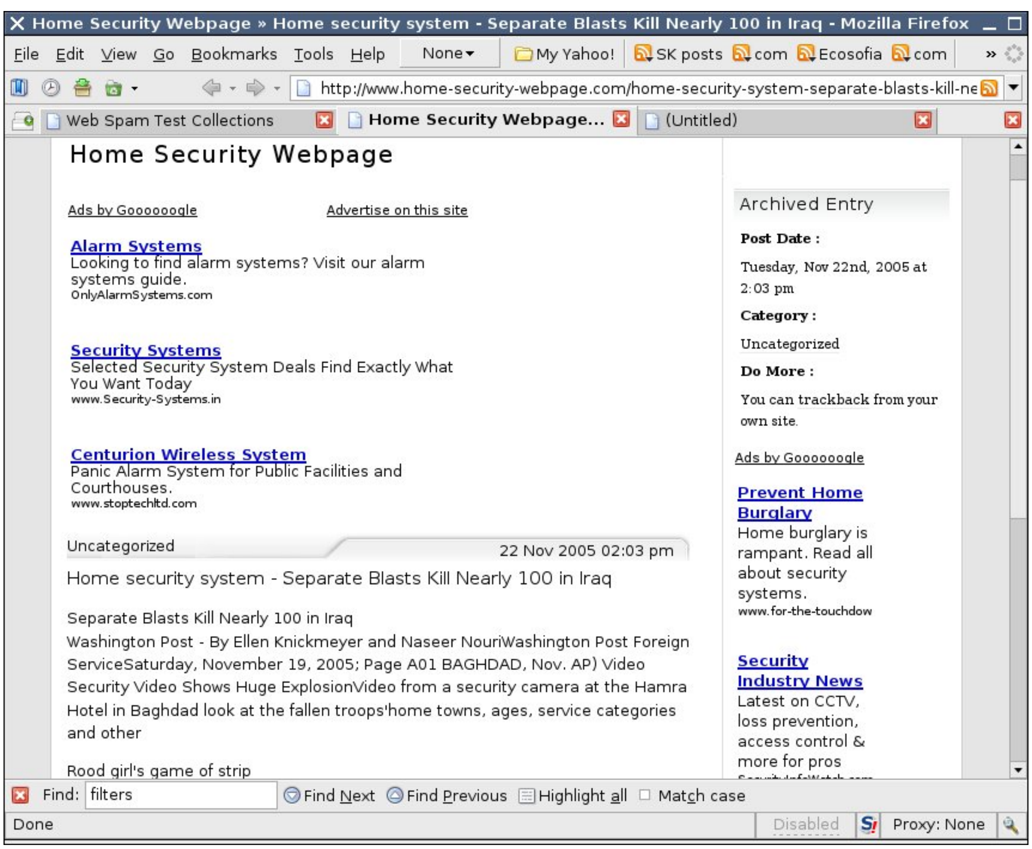
Naïve Web Spam



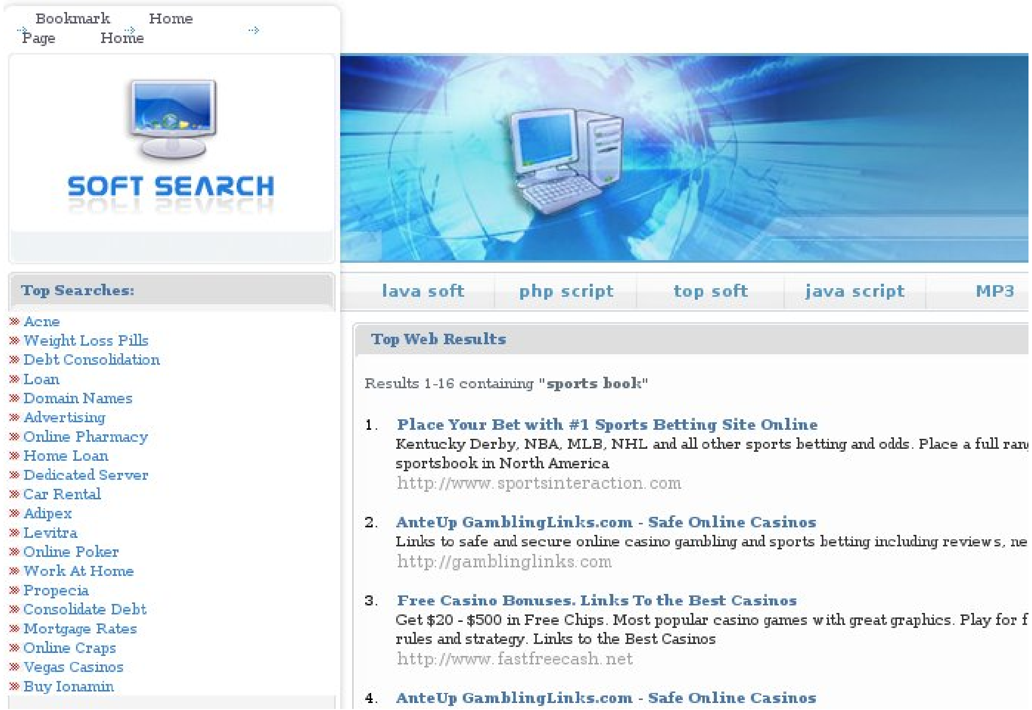
Hidden text



Made for Advertising

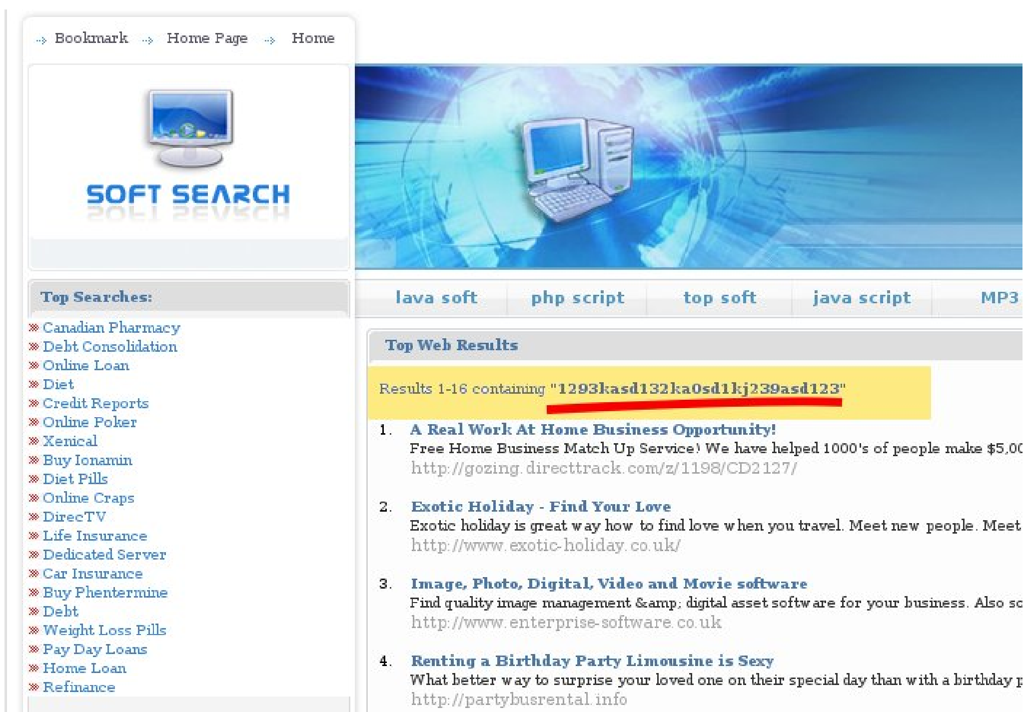


Search engine?



Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

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!

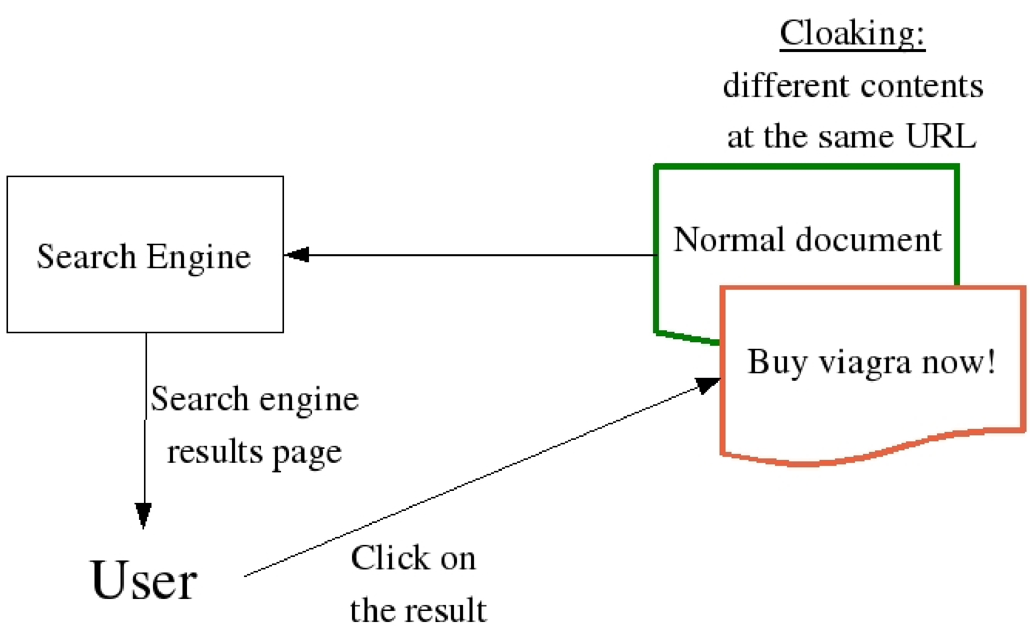
Website design, management, marketing and promotion is the key to your needs. You're one step ahead with Dry Media.

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

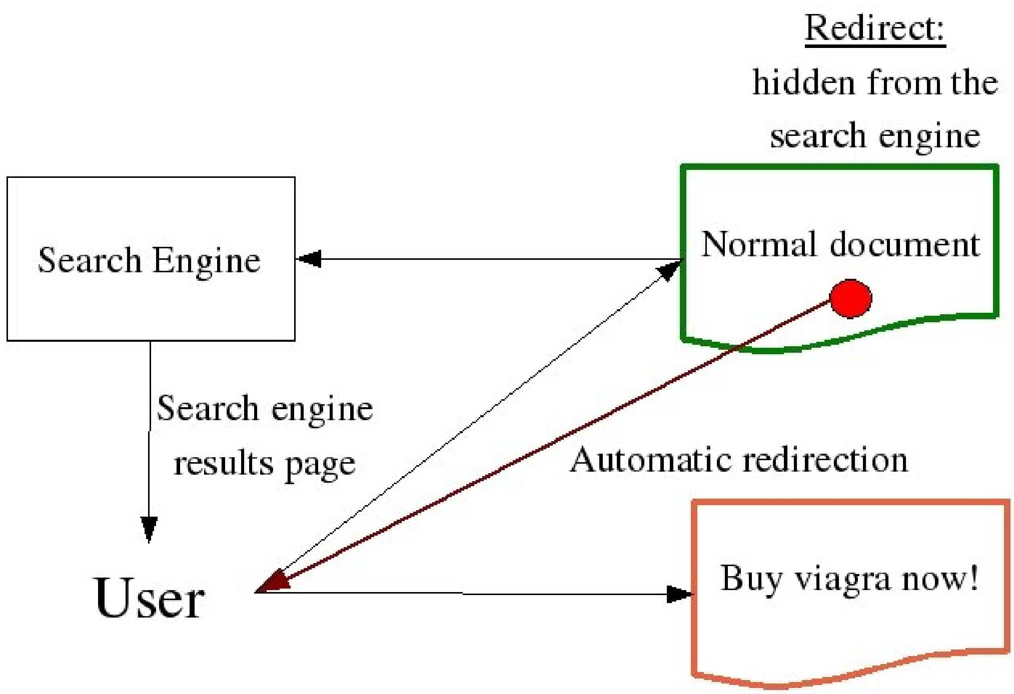
At the Website design, management, marketing and promotion web site, you'll discover an easy to use, information packed source of data on Website design, management, marketing and promotion.
Click Here to Learn More about Website design, management, marketing and promotion.

Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Cloaking



Redirection



Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

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>
```

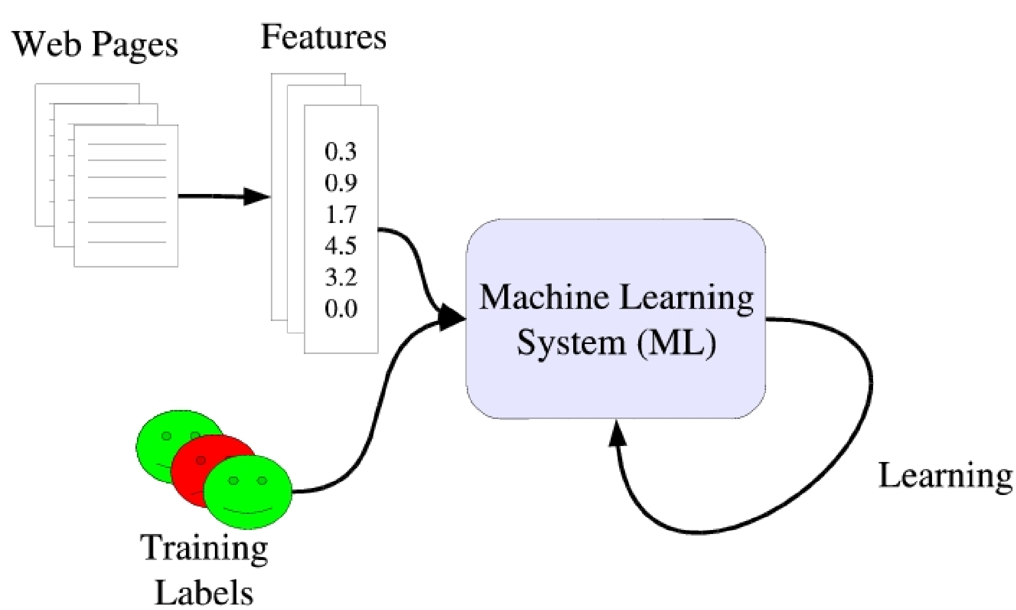
More examples: [Chellapilla and Maykov, 2007]

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

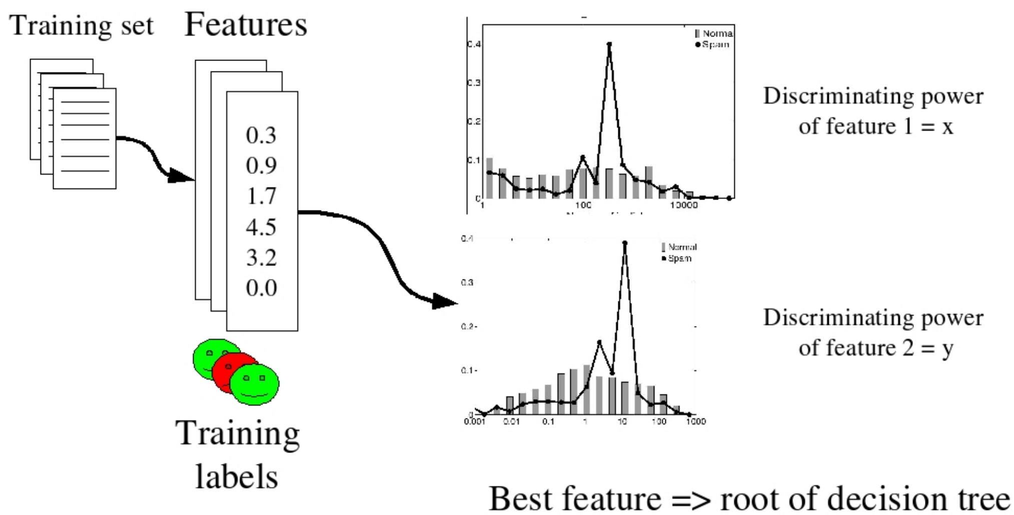
1 Web Spam
2 Web Spam Detection
3 A Reference Collection
4 Web Links
5 Topological Web Spam
6 Counting of Supporters
7 Content-based Spam detection
8 Web Topology
9 Conclusions

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

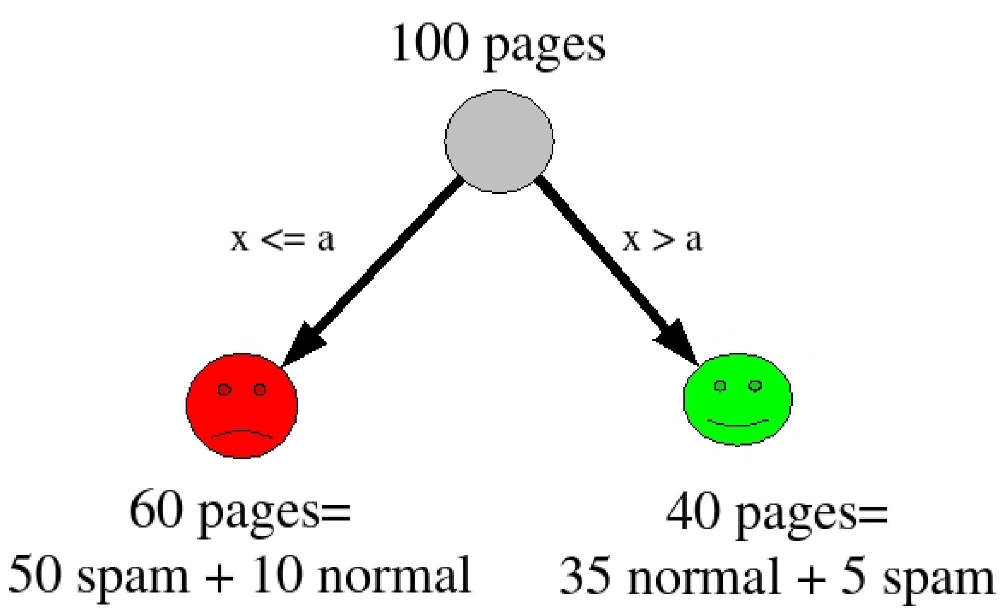
Machine Learning



Training of a Decision Tree

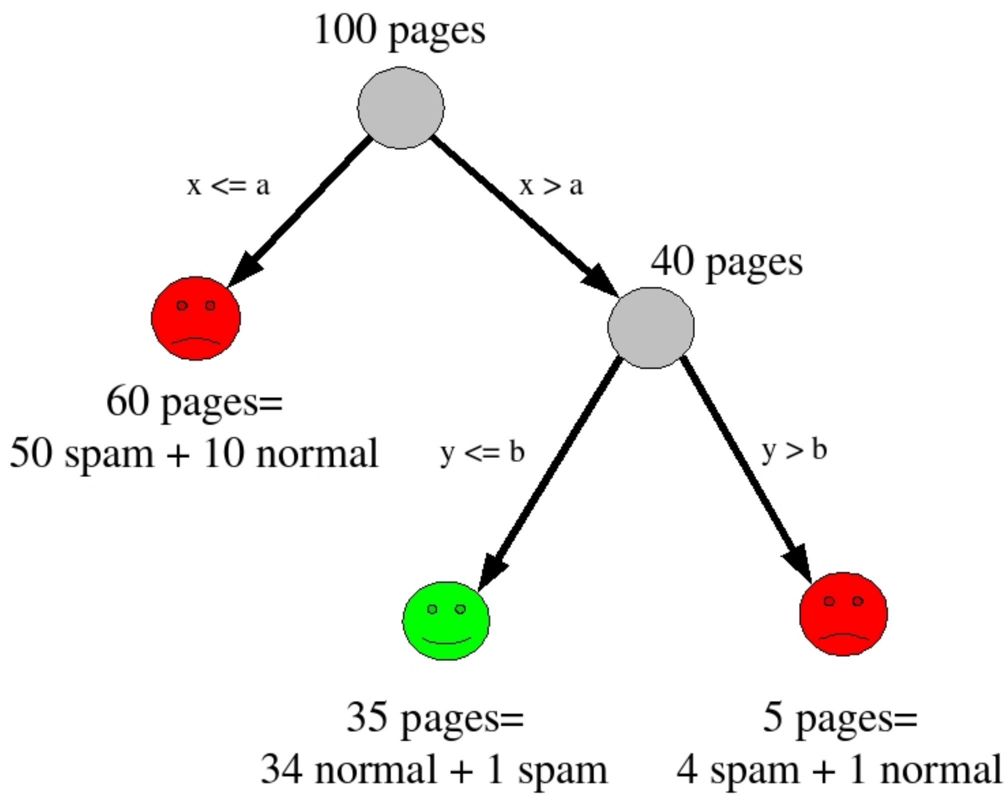


Decision Tree (error = 15%)



Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

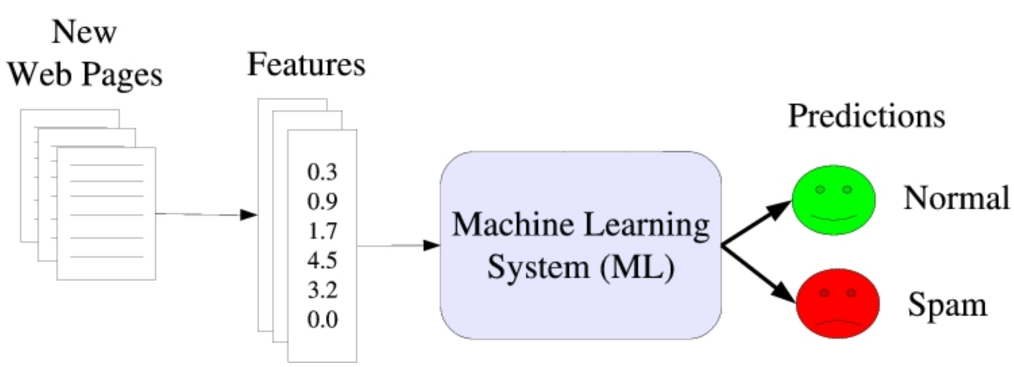
Decision Tree (error $= 15\% \rightarrow 12\%$)



Web Spam Detection
R. Baeza-Yates

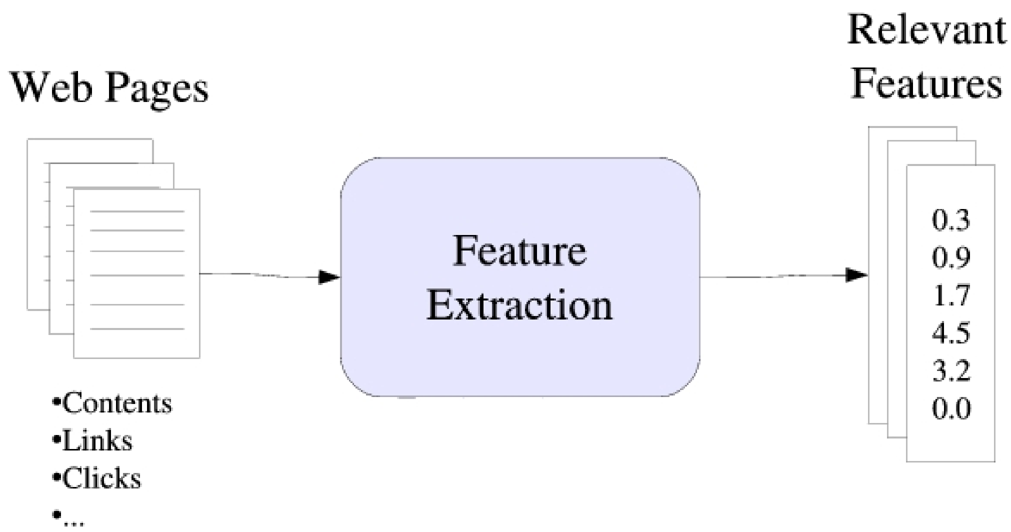
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Machine Learning (cont.)



Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Feature Extraction



Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

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

1 Web Spam
2 Web Spam Detection
3 A Reference Collection
4 Web Links
5 Topological Web Spam
6 Counting of Supporters
7 Content-based Spam detection
8 Web Topology
9 Conclusions

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

Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Assembling Process

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

Crawling of base data

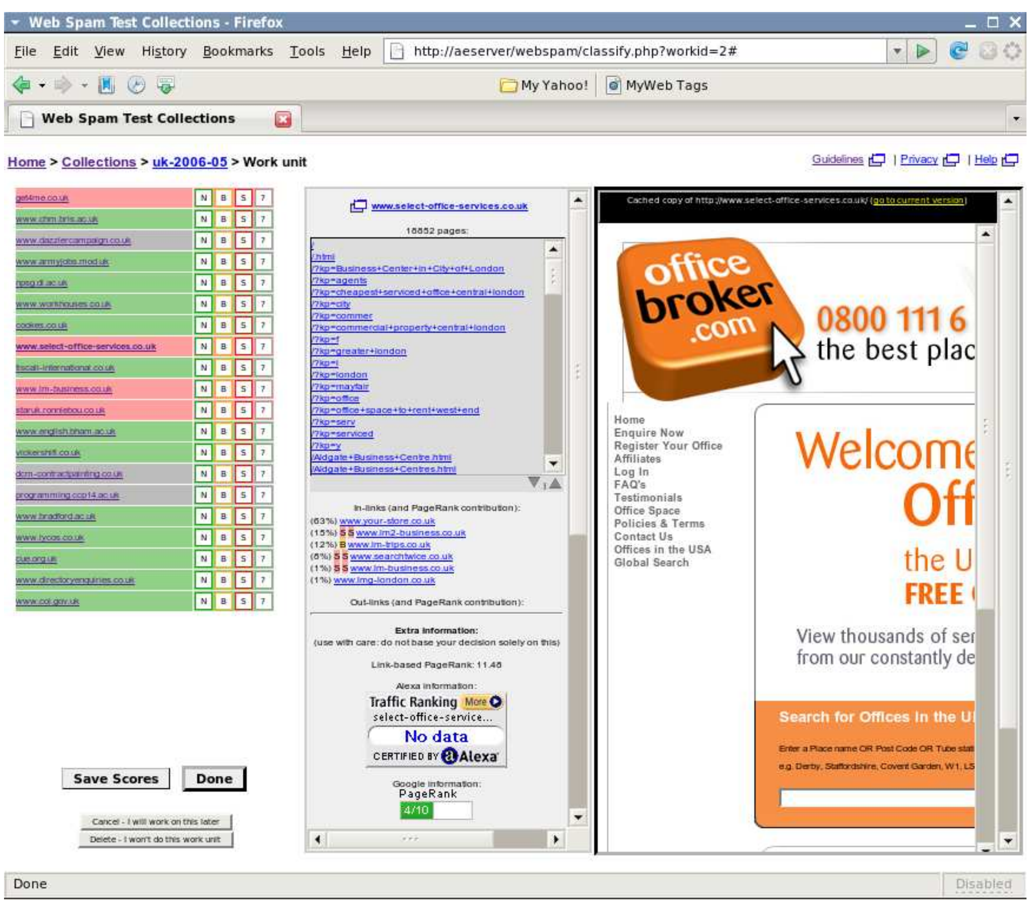
U.K. collection

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

- Large seed of about 150,000 `.uk` hosts
- 11,400 hosts
- 8 levels depth, with <=50,000 pages per host

Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Classification interface



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	infolab.hud.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2549	infoanpaydaytexas.dahanmusaprima.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	intranet.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	minal:N	alex:B	
2573	ioewebserver.ioe.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2574	ioj.ies.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	minal: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.auctionsata.co.uk	zoltan:S	sebastiano:S	
2585	itreviews.co.uk	massimo:N	brian:N	
2586	itweek.directorym.co.uk	xiaoguang:N	thiago:N	

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

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

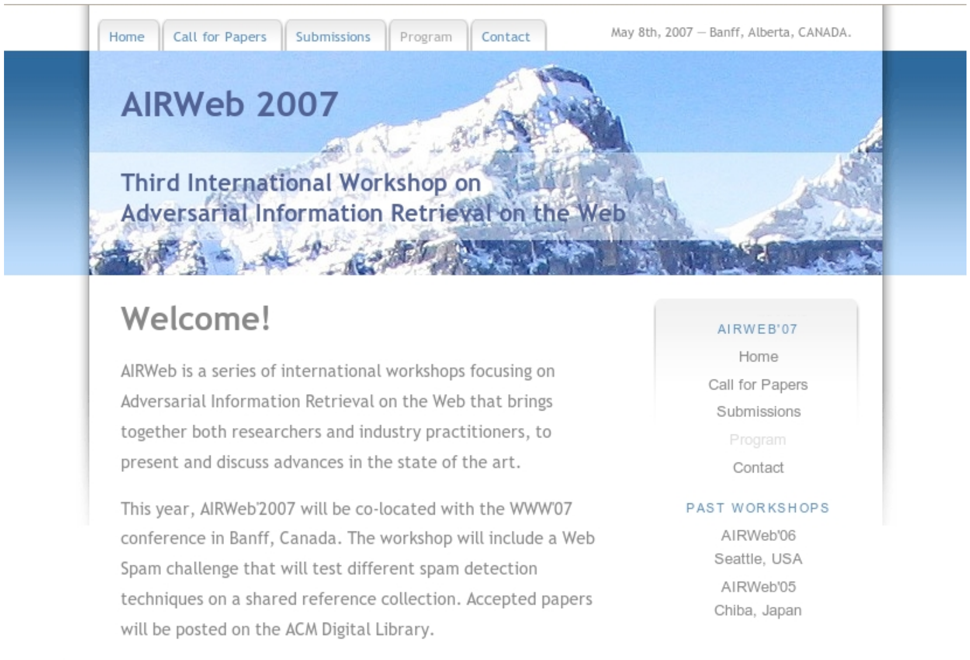
Category	Kappa	Interpretation
normal	0.62	Substantial agreement
spam	0.63	Substantial agreement
borderline	0.11	Slight agreement
global	0.56	Moderate agreement

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Result: first public Web Spam collection

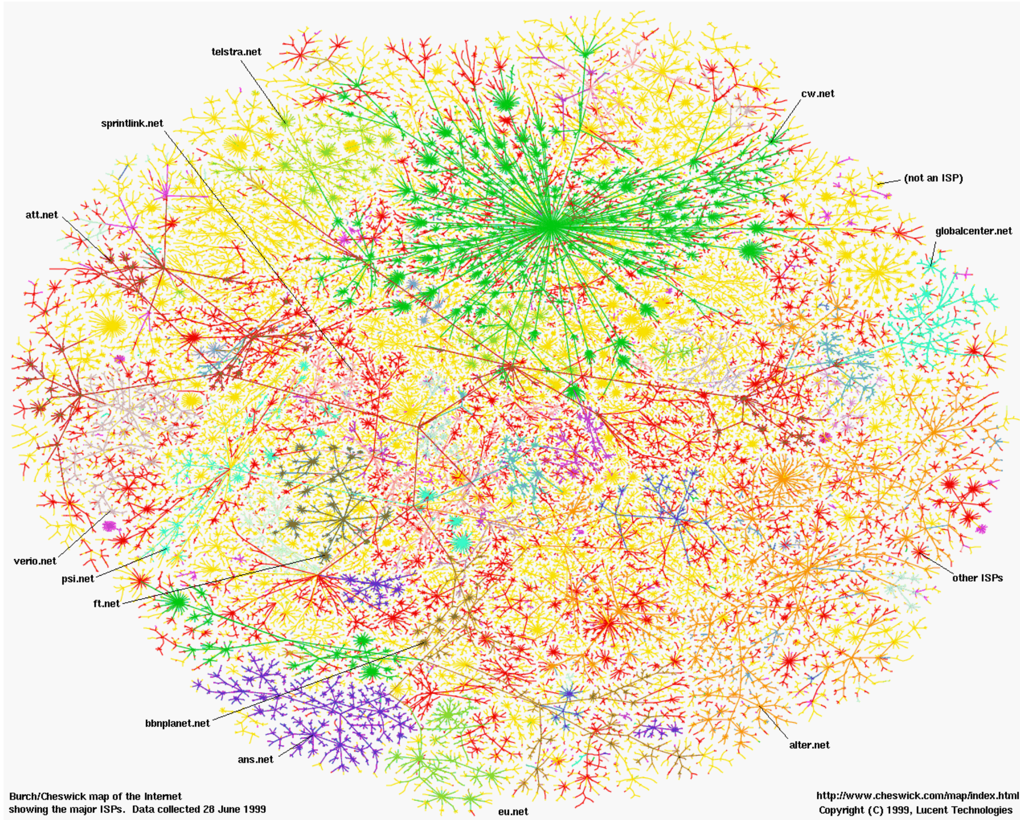
- Public spam collection
  - Labels for 6,552 hosts
  - 2,725 hosts classified by at least 2 humans
  - 3,106 automatically considered normal (`.ac.uk`, `.sch.uk`, `.gov.uk`, `.mod.uk`, `.nhs.uk` or `.police.uk`)
  - **http://www.yr-bcn.es/webspam/**
- Upcoming Web Spam challenge
  - Track I: Information retrieval + Machine learning
  - Track II: Machine learning
  - **http://webspam.lip6.fr/**
- AIRWeb 2007 Workshop (challenge results available)
  - Regular and short papers
  - Track I of the Web Spam Challenge
  - **http://airweb.cse.lehigh.edu/2007/**

AIRWeb 2007 in Banff, Canada



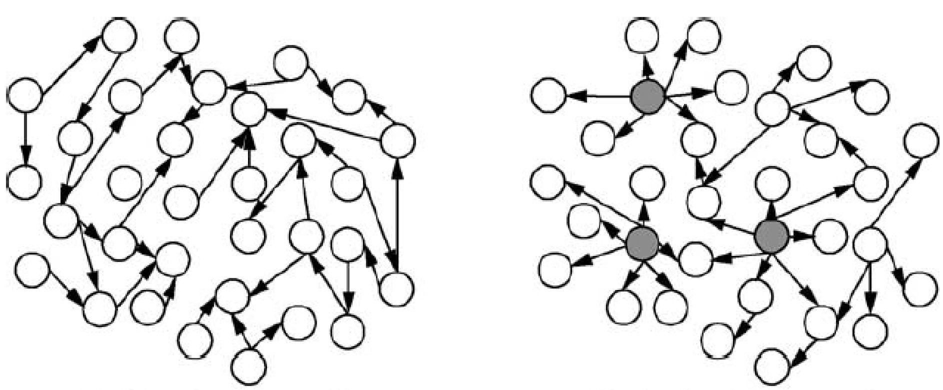
Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

1 Web Spam
2 Web Spam Detection
3 A Reference Collection
4 Web Links
5 Topological Web Spam
6 Counting of Supporters
7 Content-based Spam detection
8 Web Topology
9 Conclusions



Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Scale-free networks



(a) Random network (b) Scale-free network

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

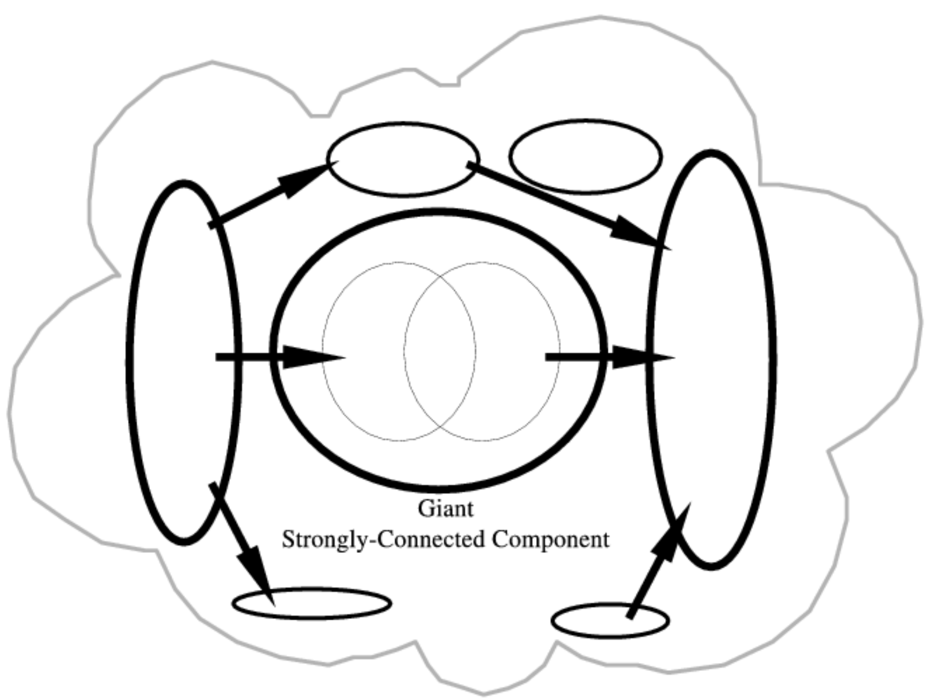
How to find meaningful patterns?

Several levels of analysis:

- Macroscopic view: overall structure
- Microscopic view: nodes
- Mesoscopic view: regions

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Macroscopic view, e.g. Bow-tie



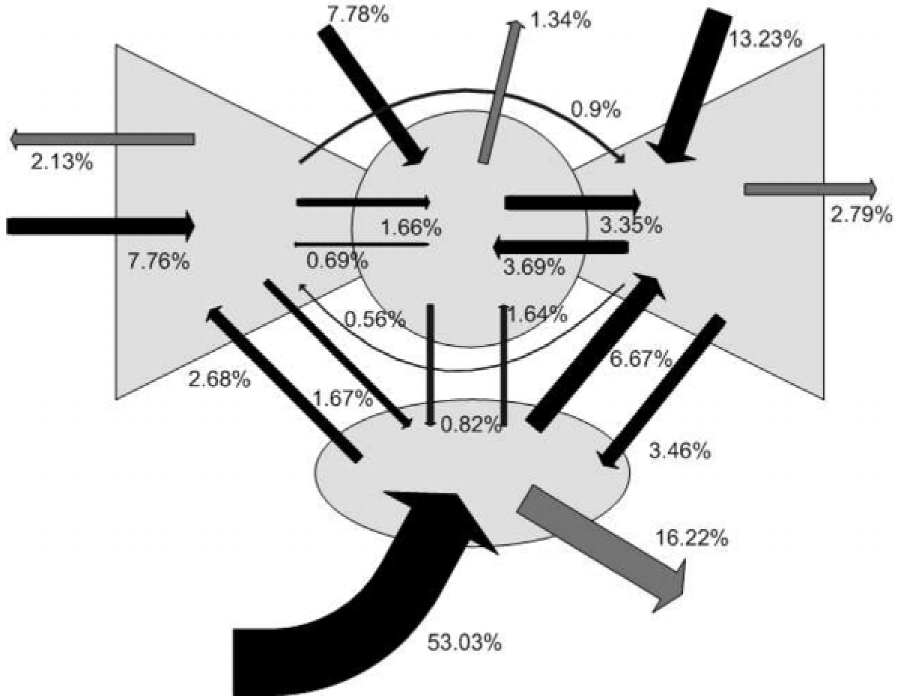
[Broder et al., 2000]

Web Spam Detection

R. Baeza-Yates

- Web Spam
- Web Spam Detection
- A Reference Collection
- Web Links
- Topological Web Spam
- Counting of Supporters
- Content-based Spam detection
- Web Topology
- Conclusions

Macroscopic view, e.g. Bow-tie, migration



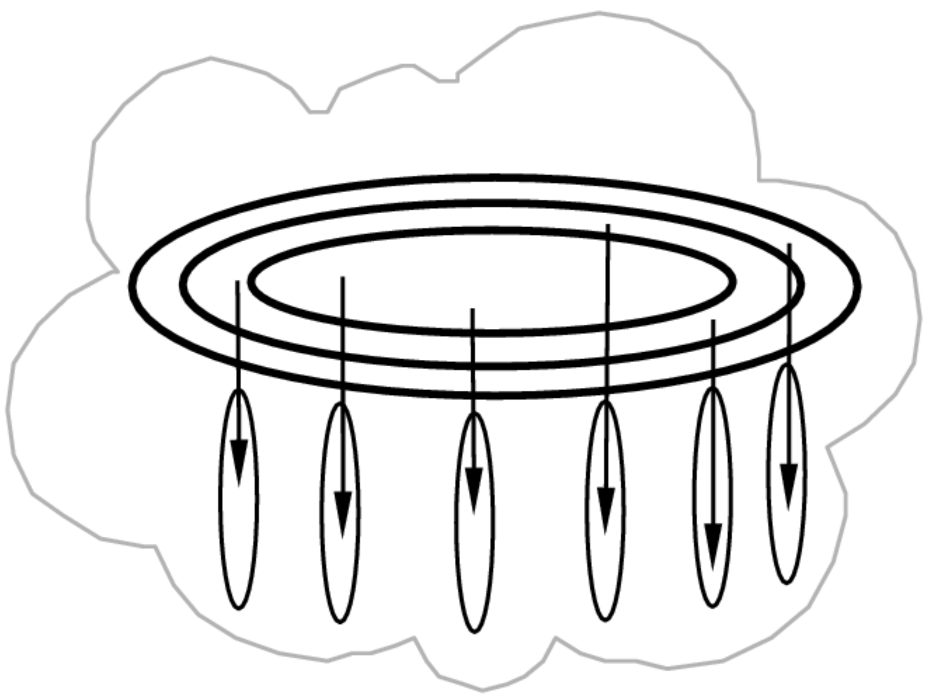
[Baeza-Yates and Poblete, 2006]

Web Spam Detection

R. Baeza-Yates

- Web Spam
- Web Spam Detection
- A Reference Collection
- Web Links
- Topological Web Spam
- Counting of Supporters
- Content-based Spam detection
- Web Topology
- Conclusions

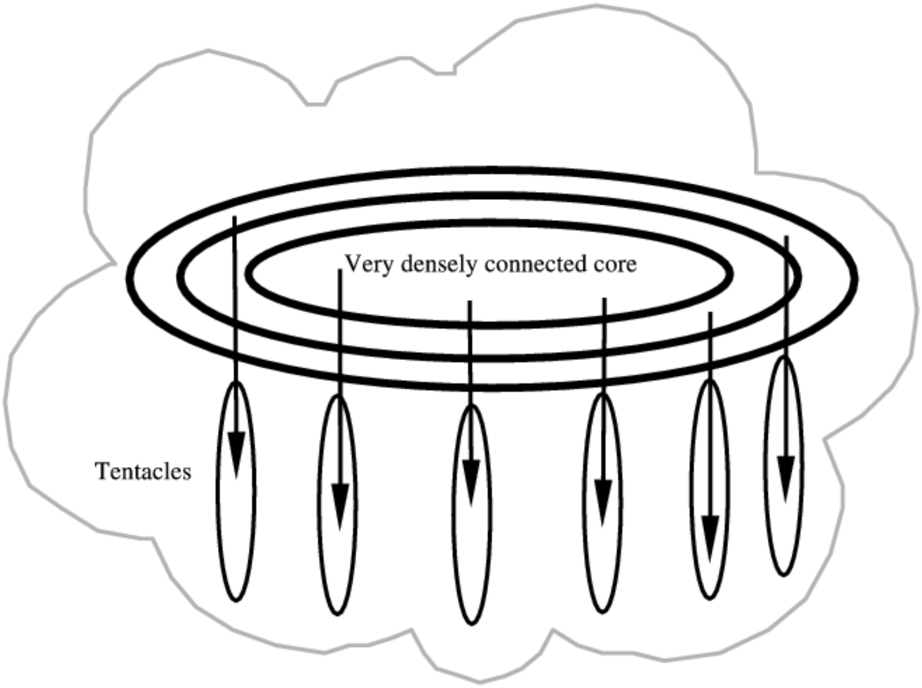
Macroscopic view, e.g. Jellyfish



[Tauro et al., 2001] - Internet Autonomous Systems (AS) Topology

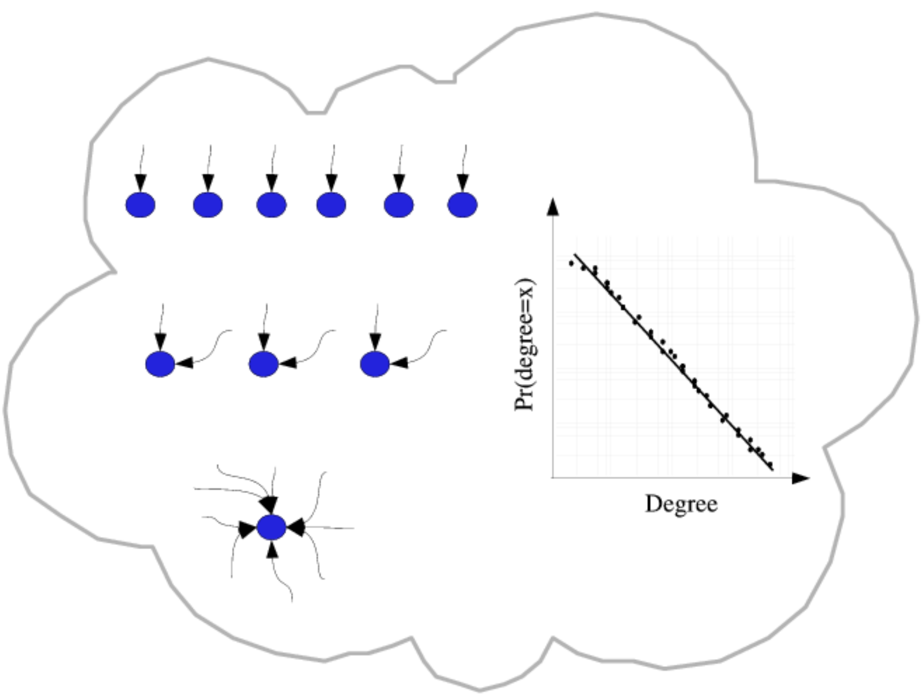
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Macroscopic view, e.g. Jellyfish



Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

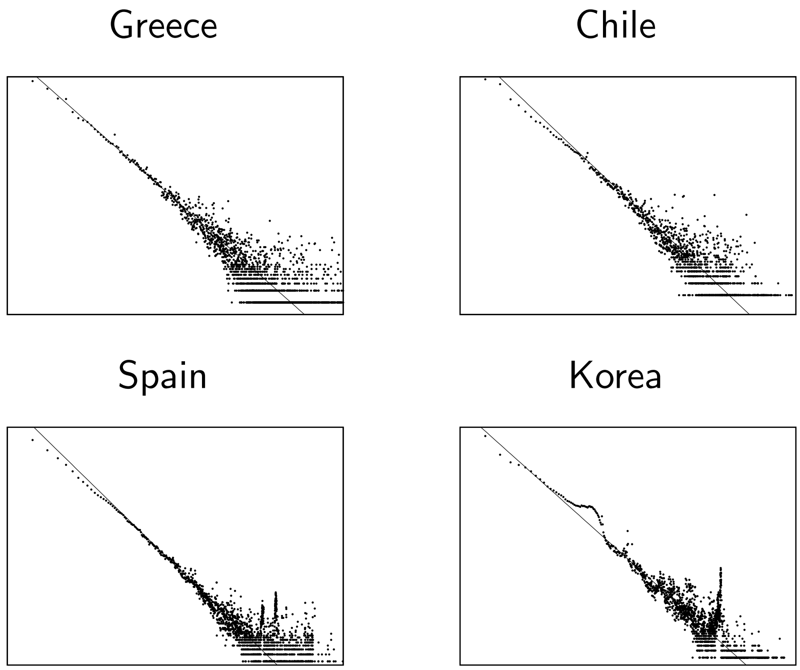
Microscopic view, e.g. Degree



[Barabási, 2002] and others

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

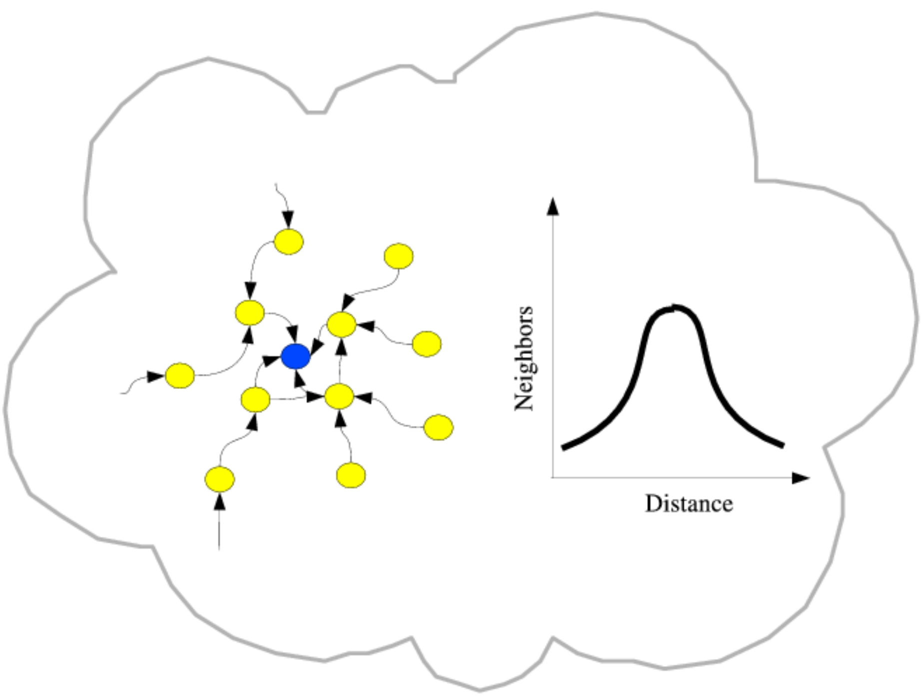
Microscopic view, e.g. Degree



[Baeza-Yates et al., 2006b] - compares this distribution in 8 countries ... guess what is the result?

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Mesoscopic view, e.g. Hop-plot

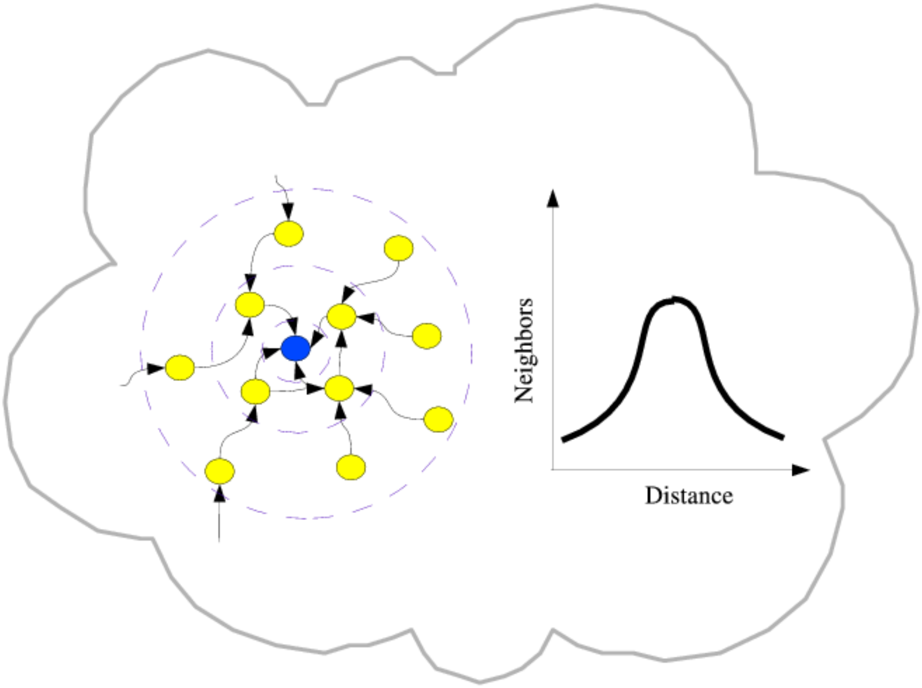


Web Spam
Detection

R. Baeza-Yates

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Mesoscopic view, e.g. Hop-plot



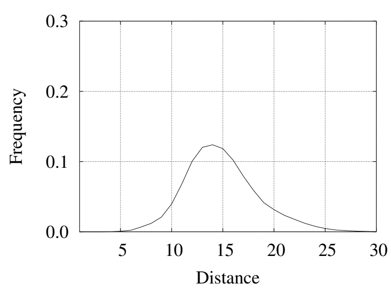
Web Spam
Detection

R. Baeza-Yates

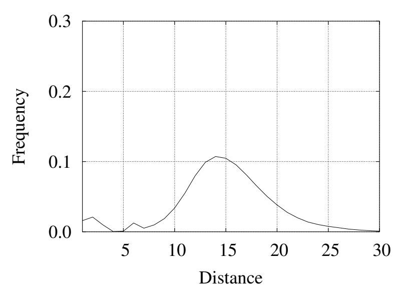
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Mesoscopic view, e.g. Hop-plot

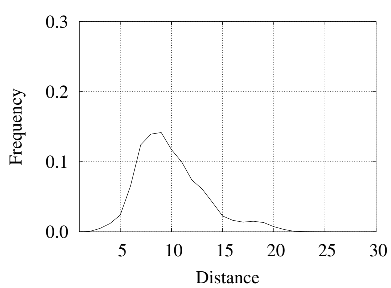
.it (40M pages)



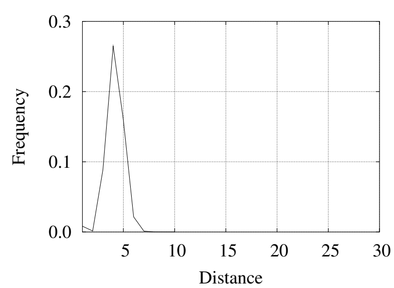
.uk (18M pages)



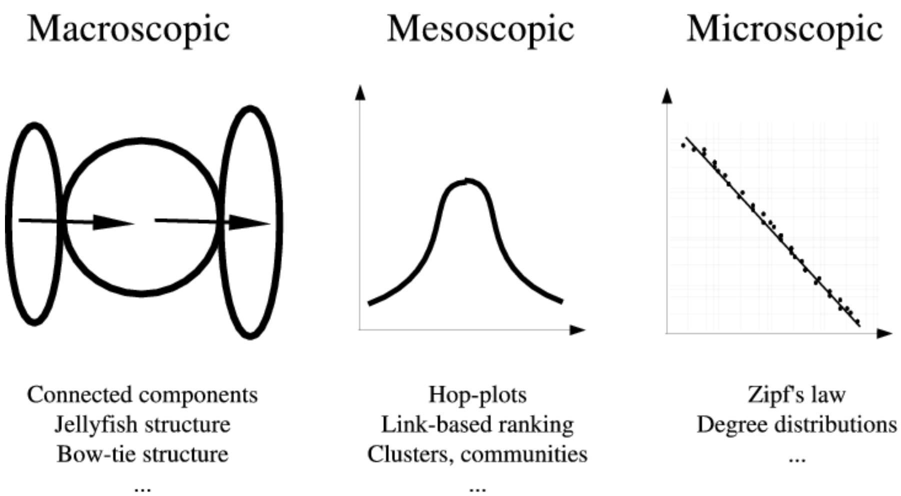
.eu.int (800K pages)



Synthetic graph (100K pages)



[Baeza-Yates et al., 2006a]



Web Spam Detection

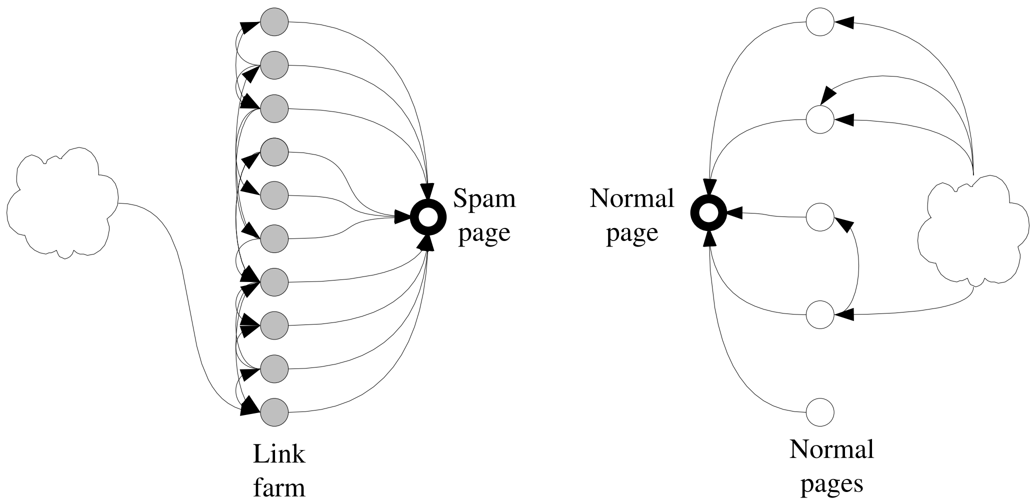
R. Baeza-Yates

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

1 Web Spam
2 Web Spam Detection
3 A Reference Collection
4 Web Links
5 Topological Web Spam
6 Counting of Supporters
7 Content-based Spam detection
8 Web Topology
9 Conclusions

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Topological spam: link farms



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

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

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**"

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Handling large graphs

For large graphs, random access is not possible.

Large graphs do not fit in main memory

Streaming model of computation

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Semi-streaming model

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

Restriction

Semi-streaming model: graph on disk

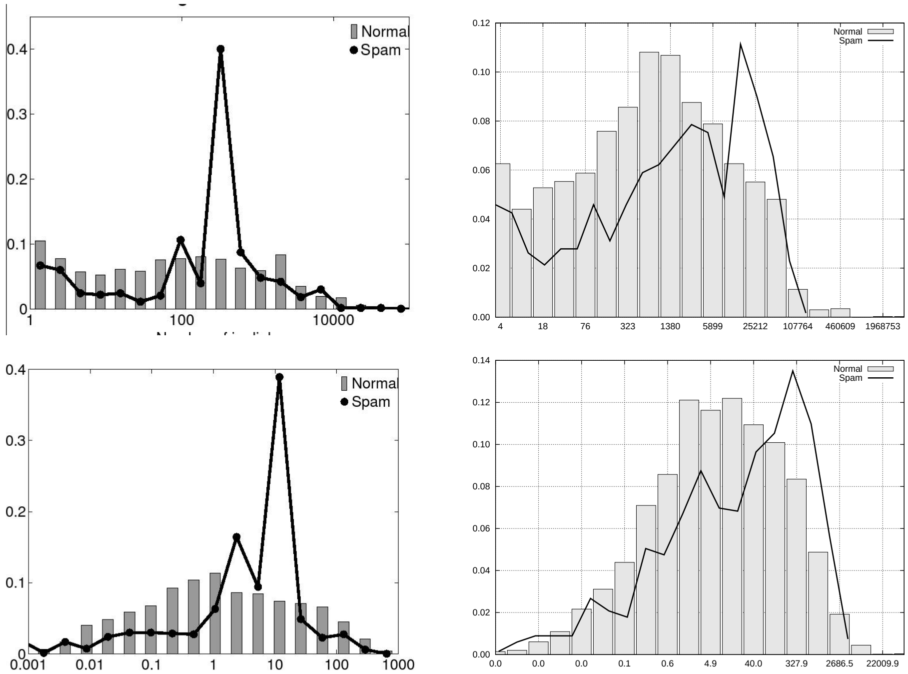
```
for node : 1 ... N do
   INITIALIZE-MEM(node)
end for
for distance : 1...d do {Iteration step}
   for src : 1 ... N do {Follow links in the graph}
      for all links from src to dest do
         COMPUTE(src,dest)
      end for
   end for
   NORMALIZE
end for
POST-PROCESS
return Something
```


Link-Based Features

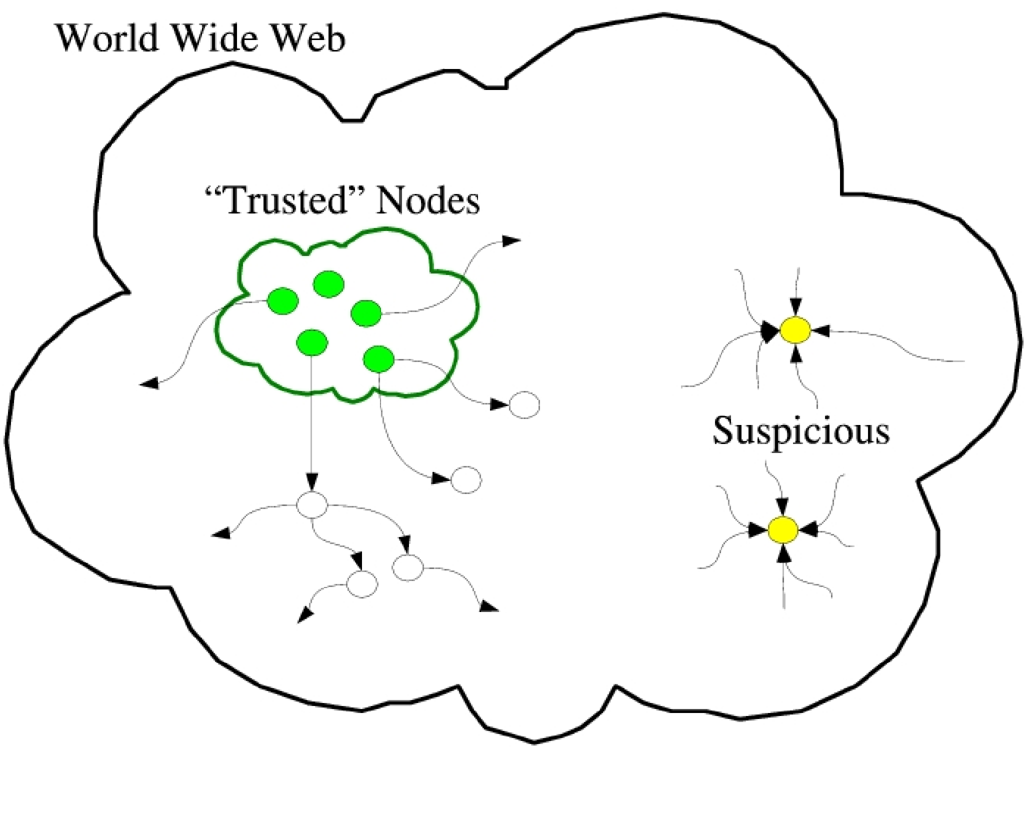
- Degree-related measures
- PageRank
- TrustRank [Gyöngyi et al., 2004]
- Truncated PageRank [Becchetti et al., 2006]
- Estimation of supporters [Becchetti et al., 2006]

140 features per host (2 pages per host)

Degree-Based



TrustRank Idea



Web Spam Detection

R. Baeza-Yates

Web Spam

Web Spam Detection

A Reference Collection

Web Links

Topological Web Spam

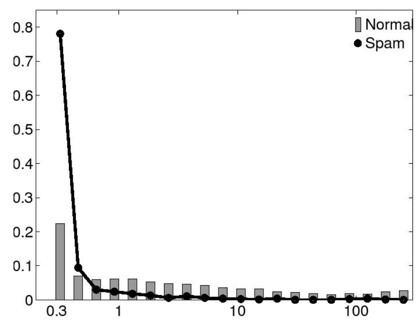
Counting of Supporters

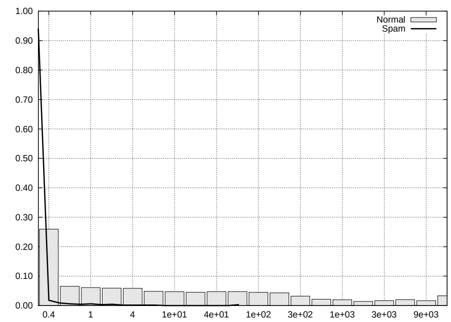
Content-based Spam detection

Web Topology

Conclusions

TrustRank / PageRank





Web Spam Detection

R. Baeza-Yates

Web Spam

Web Spam Detection

A Reference Collection

Web Links

Topological Web Spam

Counting of Supporters

Content-based Spam detection

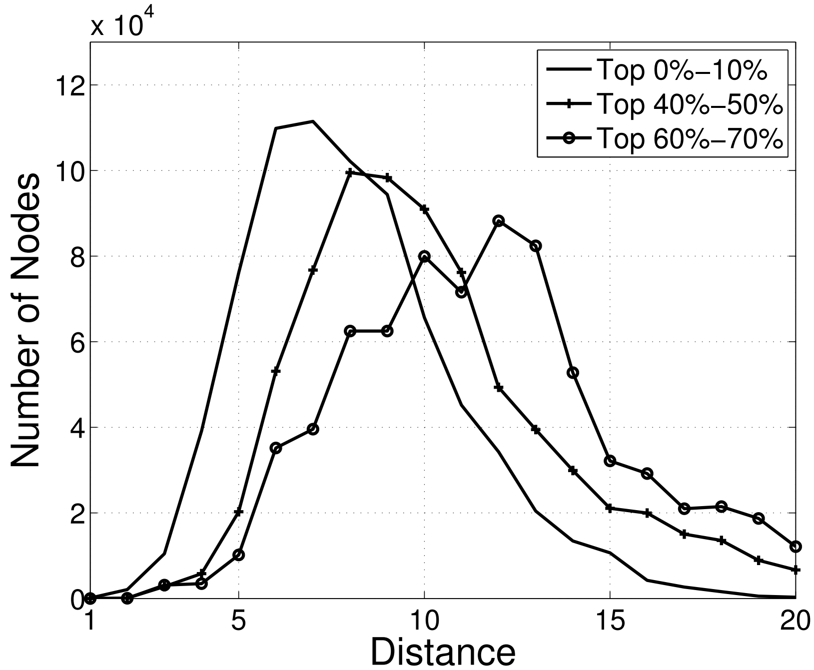
Web Topology

Conclusions

1 Web Spam
2 Web Spam Detection
3 A Reference Collection
4 Web Links
5 Topological Web Spam
6 Counting of Supporters
7 Content-based Spam detection
8 Web Topology
9 Conclusions

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

High and low-ranked pages are different

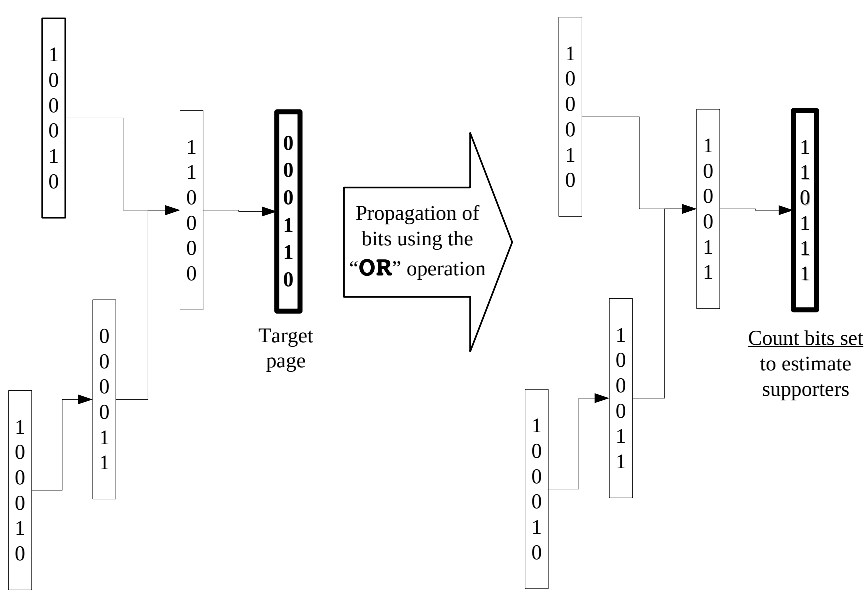


Areas below the curves are equal if we are in the same strongly-connected component

Web Spam Detection
R. Baeza-Yates

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

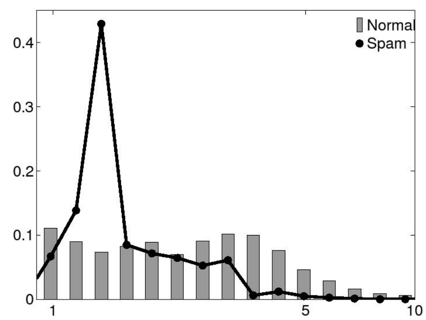
Probabilistic counting

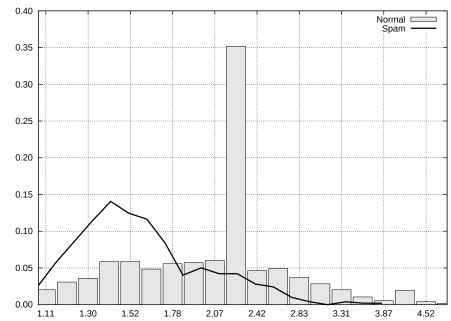


[Becchetti et al., 2006] shows an improvement of ANF algorithm [Palmer et al., 2002] based on probabilistic counting [Flajolet and Martin, 1985]

Bottleneck number

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





Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

1 Web Spam
2 Web Spam Detection
3 A Reference Collection
4 Web Links
5 Topological Web Spam
6 Counting of Supporters
7 Content-based Spam detection
8 Web Topology
9 Conclusions

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

96 features per host

Average word length

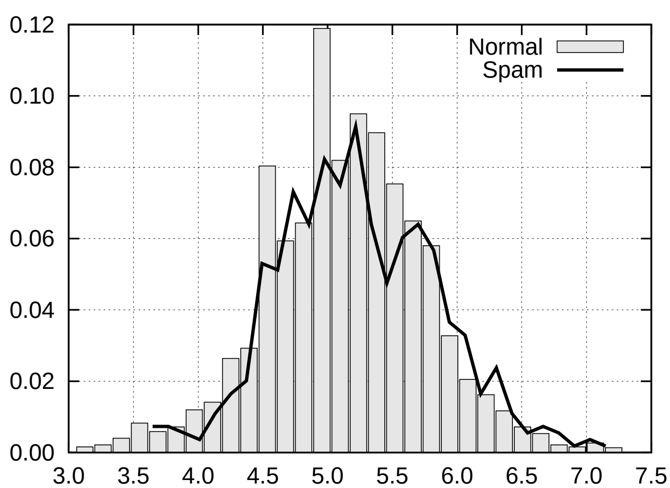


Figure: Histogram of the average word length in non-spam vs. spam pages for $k = 500$.

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Corpus precision

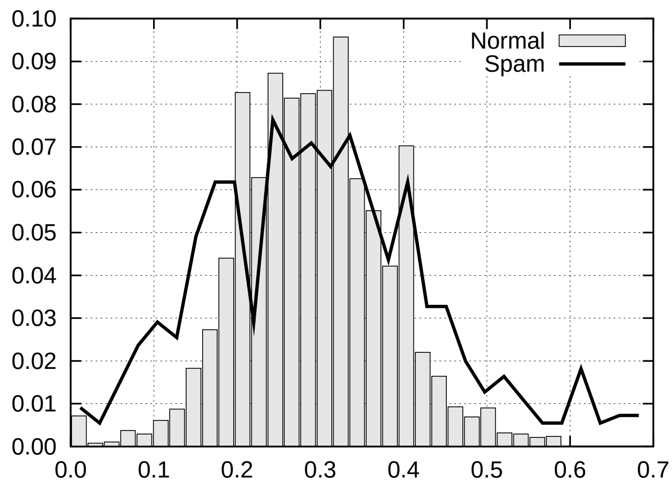


Figure: Histogram of the corpus precision in non-spam vs. spam pages.

Web Spam Detection
R. Baeza-Yates

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Query precision

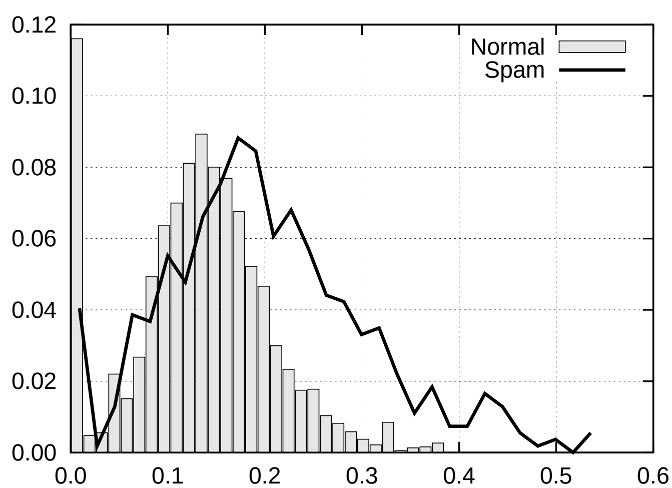


Figure: Histogram of the query precision in non-spam vs. spam pages for $k = 500$.

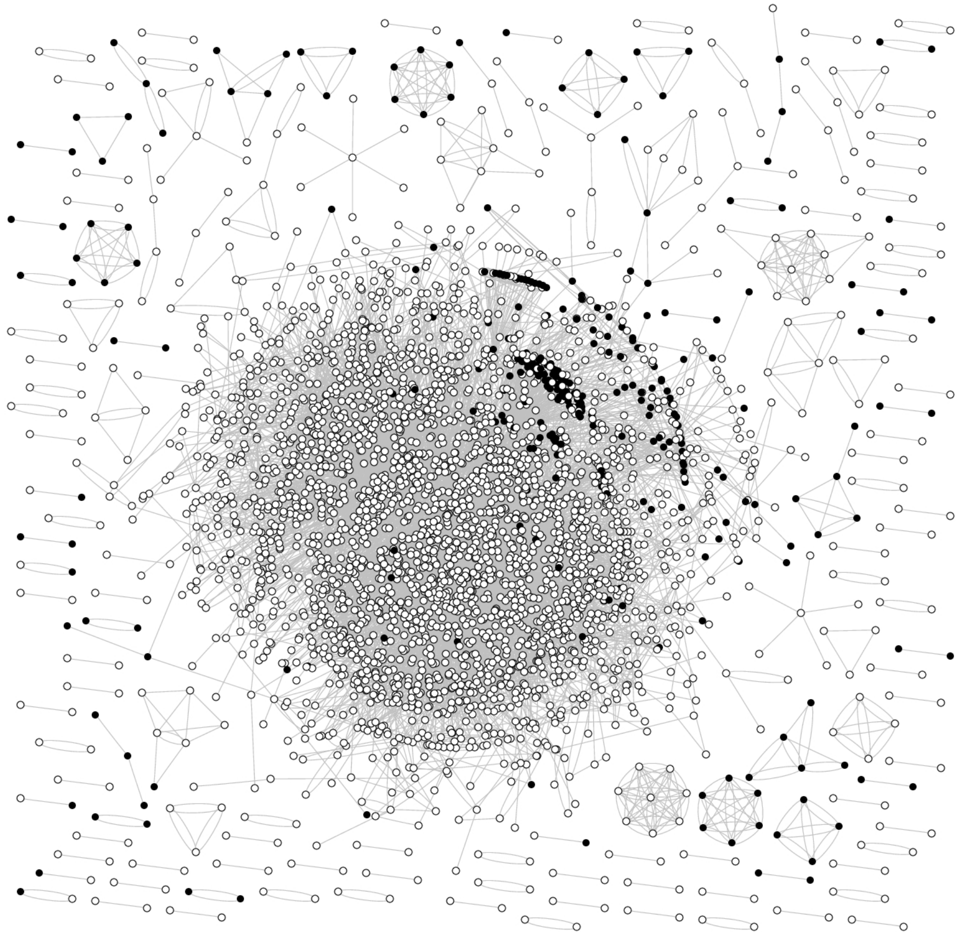
1 Web Spam
2 Web Spam Detection
3 A Reference Collection
4 Web Links
5 Topological Web Spam
6 Counting of Supporters
7 Content-based Spam detection
8 Web Topology
9 Conclusions

Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

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]



Web Spam Detection

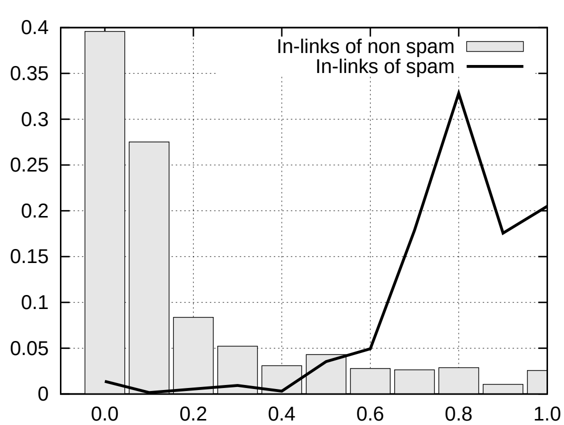
R. Baeza-Yates

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Topological dependencies: in-links

Histogram of fraction of spam hosts in the in-links

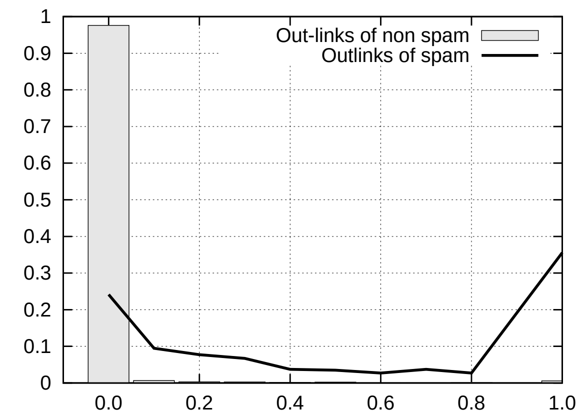
- 0 = no in-link comes from spam hosts
- 1 = all of the in-links come from spam hosts



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



Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Idea 1: Clustering

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

	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

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Idea 2: Propagate the label

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

	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

Web Spam Detection
R. Baeza-Yates

Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Idea 3: Stacked graphical learning

Classify, then add the average predicted “spamicity” of neighbors as a new feature for each node, then classify again[Cohen and Kou, 2006]

	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

Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

1 Web Spam
2 Web Spam Detection
3 A Reference Collection
4 Web Links
5 Topological Web Spam
6 Counting of Supporters
7 Content-based Spam detection
8 Web Topology
9 Conclusions

Concluding remarks

- ☑ The UK-2006-05 dataset is "harder" than previous datasets
- ☑ Considering content-based and link-based attributes improves the accuracy
- ☑ Considering the dependencies improves the accuracy

Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Thank you!

Baeza-Yates, R., Boldi, P., and Castillo, C. (2006a).
Generalizing pagerank: Damping functions for link-based ranking algorithms.
In *Proceedings of ACM SIGIR*, pages 308–315, Seattle, Washington, USA. ACM Press.

Baeza-Yates, R., Castillo, C., and Efthimiadis, E. (2006b).
Characterization of national web domains.
To appear in ACM TOIT.

Baeza-Yates, R. and Poblete, B. (2006).
Dynamics of the chilean web structure.
Comput. Networks, 50(10):1464–1473.

Barabási, A.-L. (2002).
Linked: The New Science of Networks.
Perseus Books Group.

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.

Web Spam Detection
R. Baeza-Yates
Web Spam
Web Spam Detection
A Reference Collection
Web Links
Topological Web Spam
Counting of Supporters
Content-based Spam detection
Web Topology
Conclusions

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000).
Graph structure in the web: Experiments and models.
In *Proceedings of the Ninth Conference on World Wide Web*, pages 309–320, Amsterdam, Netherlands. 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.

Tauro, L., Palmer, C., Siganos, G., and Faloutsos, M. (2001).
A simple conceptual model for the internet topology.
In *Global Internet*, San Antonio, Texas, USA. IEEE CS Press.