
Improving Web Spam Classifiers Using Link Structure

Qingqing Gan and Torsten Suel
Polytechnic University
Brooklyn, New York

Outline

- Spam techniques
- Spam detection techniques
- Baseline classifier
- Neighborhood structure of spam
- Improvements to the baseline classifier
- Summary

Web Spam – Any aggressive attempt to obtain a higher-than-deserved position by manipulating search engine ranking methods.

Spam Techniques

Content Spam

- Body spam, title spam, meta-tag spam and anchor text spam
- Keyword stuffing
- Hidden text
- Cloaking
- Redirection

Link spam

- Page content copy
- Infiltrating a web directory
- Link exchange
- Blog spam

Spam Detection Techniques

- Learning-based classification
 - I. Dorst and T. Scheffer. *Thwarting the nigritude ultramarine: Learning to identify link spam.*
 - D. Fetterly, M. Manasse, and M. Najork. *Spam, dam spam, and statistics: using statistical analysis to locate spam web pages.*
 - A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. *Detecting spam web pages through content analysis.*
- Trust and distrust propagation
 - B. Wu and B. Davison. *Identifying link farm spam pages.*
 - Z. Gyongyi, H. Garcia-Moline, and J. Pedersen. *Combating web spam with trustrank.*
 - M. Sobek. *PR0 – Google's PageRank 0 penalty, 2002.*
 - B. Wu, V. Goel, and B. Davison. *Propagating trust and distrust to demote Web spam.*
- Neighborhood structure and similarity
 - A. Benczur, K. C. T., and Sarlos. *Link-based similarity search to fight web spam.*
 - S. Charkrabarti, B. Dom, and P. Indyk. *Enhanced hypertext categorization using hyperlinks.*
 - B. Davison. *Recognizing nepotistic links on the web.*
 - L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. *Link-based characterization and detection of Web Spam.*
 - C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. *Know your neighbors: Web spam detection using the web topology.* Technical report.
 - E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. So. *The connectivity sonar: Detecting site functionality by structural patterns.*

Dataset

- Swiss ch top-level domain crawled in 2005 using PolyBot web crawler
- 12 million pages located on 239,272 hosts, connected by 234 million links
- A site-level web graph
- Sampled and manually examined 1000 sites (500 spam and 500 non-spam ones)

Basic Classifier

Content features:

- Number of words in a page
 - Average length of words in a page
 - Fraction of words drawn from globally popular words
 - Fraction of globally popular words used in a page
 - Fraction of visible content
 - Number of words in the page title
 - Amount of anchor text in a page
 - Compression rate of the page
- (A. Ntoulas, et al. WWW2006)

Three other features we added:

- Number of hosts in the domain
- Ratio of pages in this host to pages in this domain
- Number of hosts on the same IP address.

Link features:

- Percentage of pages in most populated level
 - Top level page expansion ratio
 - In-links per page
 - Out-links per page
 - Out-links per in-link
 - Top-level in-link portion
 - Out-links per leaf page
 - Average level of in-links
 - Average level of out-links
 - Percentage of in-links to most popular level
 - Percentage of out-links from most emitting level
 - Cross-links per page
 - Top-level interval in-links per page on this site
 - Average level of page in this site
- (E.Amitay, et al. Hypertext2003)

Results for Basic Classifier

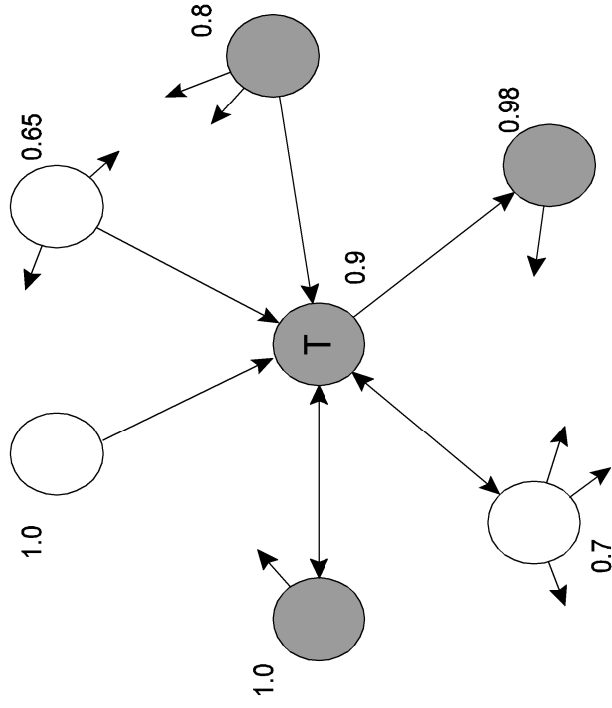
	Precision	Recall	F-measure
Spam	0.897	0.812	0.852
Non-spam	0.882	0.925	0.903

Decision Tree C4.5

	Precision	Recall	F-measure
Spam	0.879	0.812	0.844
Non-spam	0.863	0.913	0.887

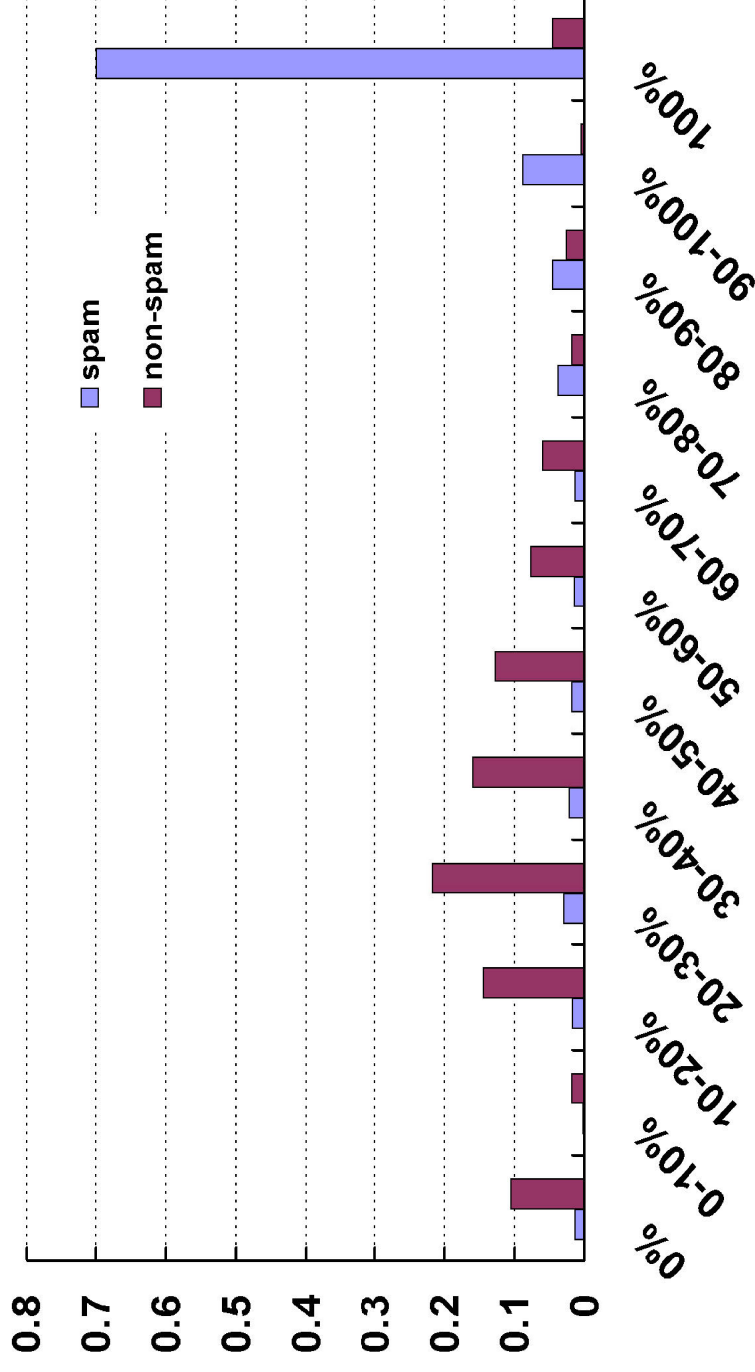
Support Vector Machine

Neighborhood Structure of Spam



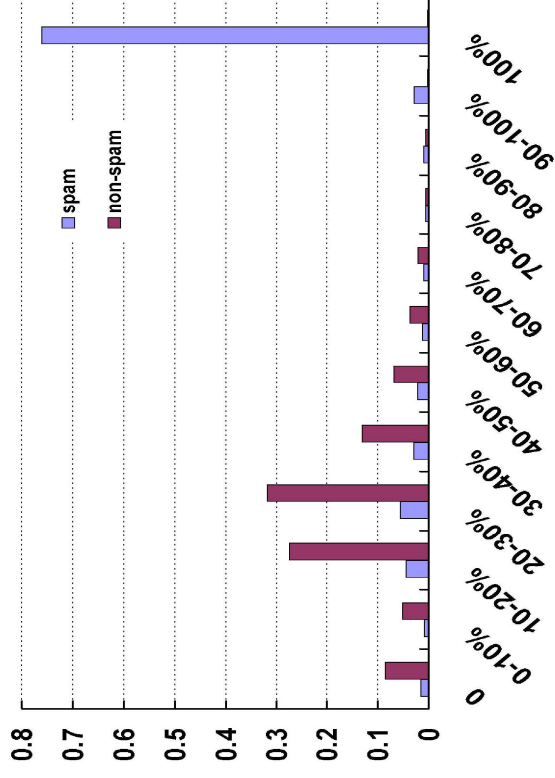
- Incoming spam distribution
- Outgoing spam distribution
- Weighted incoming spam distribution

Neighborhood Structure of Spam

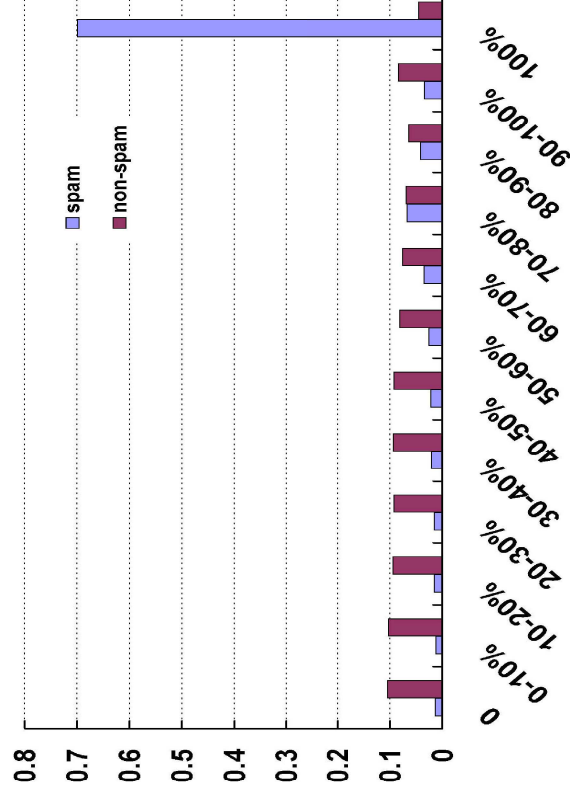


In-link spam distribution for spam and non-spam sites

Neighborhood Structure of Spam (Cont.)



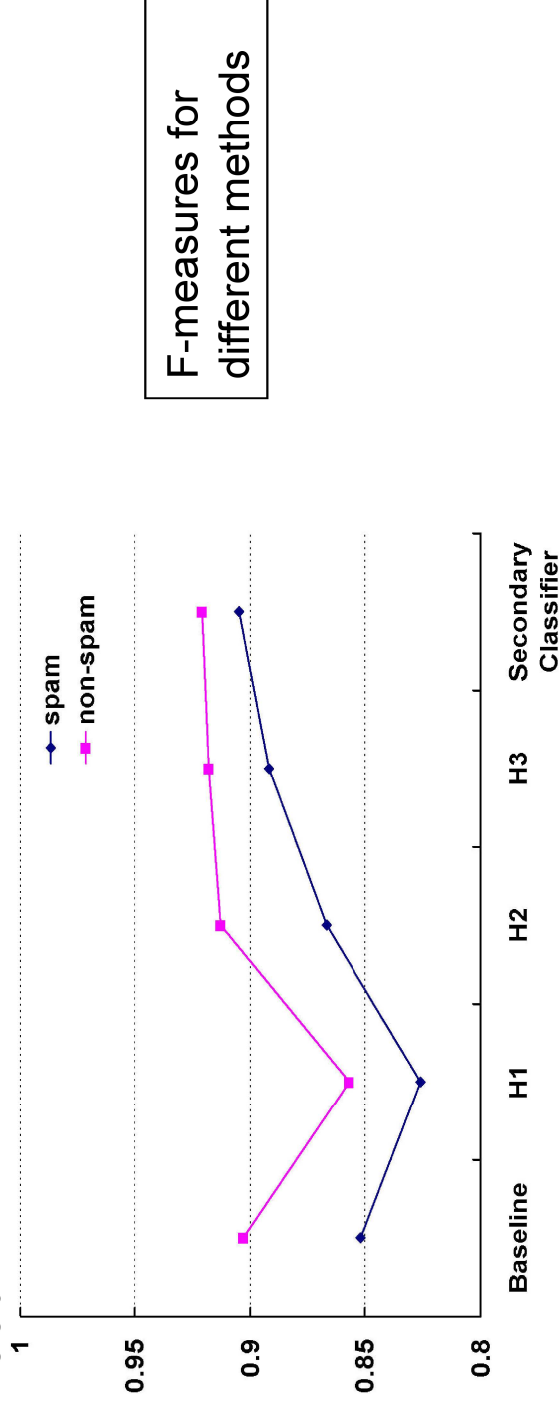
Out-link spam distribution for
spam and non-spam sites



Weighted in-link spam distribution for
spam and non-spam sites

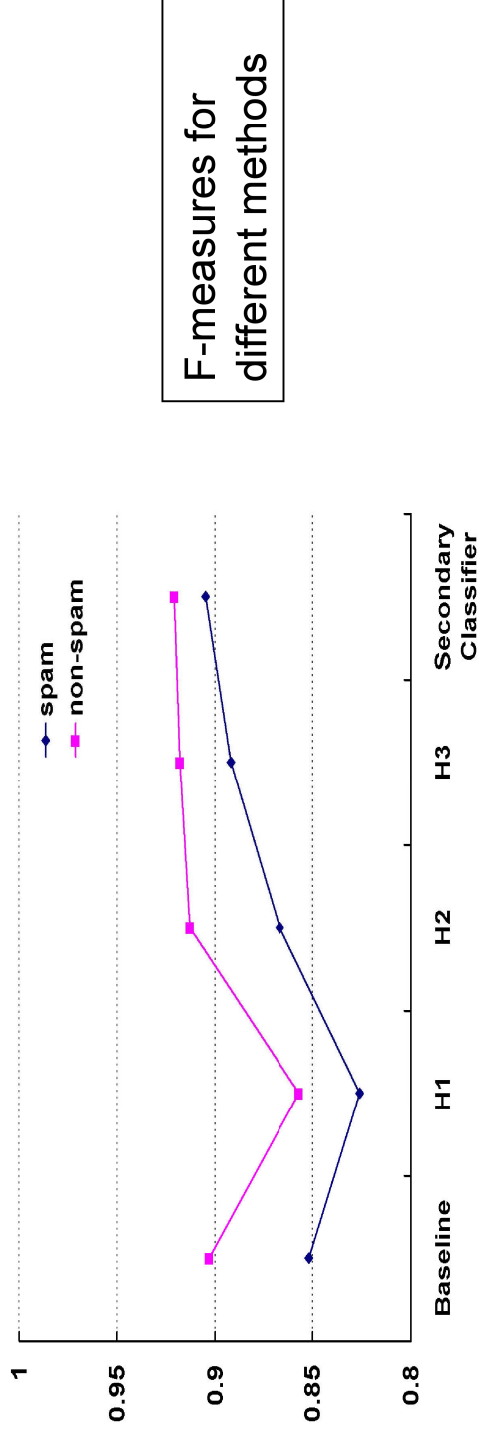
Improving the Basic Classifier - Relabeling

- H1: Relabeling according to the fraction X of spam sites in the total incoming neighborhood.
- H2: Relabeling according to the fraction of spam in the weighted incoming neighborhood
- H3: Relabeling according to the fraction of spam in the outgoing neighborhood



Improving the Basic Classifier - Secondary Classifier

- The label by the basic classifier
- The confidence score associated with the feature above
- The percentage of incoming links from spam sites
- The percentage of outgoing links pointing to spam sites
- The fraction of weighted spam in the incoming neighbors
- The fraction of weighted spam in the outgoing neighbors
- The percentage of weighted incoming spam.



Summary

- Our results show visible improvements by adding a secondary heuristic or secondary classifier
- The grey area between ethical search engine optimization and unethical spam
- In our case, a few major spam farms are responsible for most of the spam
- Should training sets be selected at random?
- Granularity of spam detection (page/site/domain/farm)
- Interactive spam detection