Combating Web Spam with TrustRank

Zoltán Gyöngyi
Hector Garcia-Molina
Jan Pedersen
“Kaiser pharmacy”

**kaiser pharmacy online**

kaiser pharmacy online. Stop surfing! Only here you can get info about kaiser pharmacy online. Click here now to read info about...

careers.ducsonsystems.net/ kaiser-pharmacy-online/kaiser-pharmacy-online.html - 2k - Cached - Similar pages

**pretty lady at kaiser pharmacy - m4w**

... pretty lady at kaiser pharmacy - m4w. Reply to: anon-37941915@craigslist.org Date: 2004-07-30, 2:03PM PDT Seeing you really brightened up my day. ... www.craigslist.org/pen/ms/37941915.html - Similar pages

**The Beautiful Pharmacy Girls at Kaiser**

... So to all of the women that work in the Kaiser Pleasanton pharmacy, I'd just like to let you know that the guy with the itchy nuts thinks all of you are ... www.craigslist.org/about/best/sfo/2705982.html - Similar pages

[ More results from www.craigslist.org ]

**Kaiser Permanente**

... In addition to outpatient, inpatient, and ambulatory care, Kaiser Permanente maintains specialized pharmacy services, automated refill pharmacies, regional ... www.kaiserpermanentejobs.org/ workinghere/specialty-pharmacy.asp - 25k - Cached - Similar pages

**Kaiser pharmacy**

KAISER PHARMACY. Education. help with debt. purchase phentermine. Consulting. Venetian Casino. extended warranty. creatine. Free Poker Games. Phone System. ...

www.virtuell-bilden.de/pharmacies--kaiser-pharmacy/ - 30k - Cached - Similar pages

**Getting Started at Kaiser Permanente**

... As a member, you may transfer a prescription from a non-Kaiser Permanente pharmacy to any of our pharmacies. Be ready to tell the ...

www.kaisersantaclara.org/choose/newmember.html - 13k - Cached - Similar pages
“Kaiser pharmacy” (CONT’D)
Roadmap

- Web spam
  - Definition, techniques
  - Observation
  - Objective

- TrustRank
  - Idea, related work
  - Score propagation
  - Algorithm

- Experiments
Roadmap

- Web spam
  - Definition, techniques
  - Observation
  - Objective

- TrustRank
  - Idea, related work
  - Score propagation
  - Algorithm

- Experiments
**Web Spam / Definition**

Spamming = misleading search engines to obtain higher-than-deserved ranking

![Diagram showing techniques of spamming and hiding](image-url)
Web Spam / Observation

Approximate isolation of good pages:
good pages seldom point to spam
Web Spam / Objective

Separate good pages from spam pages

What?
Assign high scores to very good pages

How?
Propagate scores from known good pages

When?
Use results in ranking
Roadmap

- Web spam
  - Definition, techniques
  - Observation
  - Objective

- TrustRank
  - Idea, related work
  - Score propagation
  - Algorithm

- Experiments
TrustRank / Idea

\[ \text{TrustRank} = \text{Selection} + \text{Propagation} \]

- Our approach: spam demotion
- Related work: spam detection
  - Text $\rightarrow$ machine learning
  - Link $\rightarrow$ graph clustering
TrustRank / Propagation

Simple Score Propagation
TrustRank / Propagation (CONT’D)

Score Dampening

1 -> 2
2 -> 1, 3
3 -> 1, 6
4 -> 2
5 -> 4, 6
7 -> 5

Scores:
1: ?
2: .80
3: .80
4: ?
5: .51
6: .64
7: 0
Score Splitting

TrustRank / Propagation (CONT’D)
1. Assess seed-desirability of pages

2. Select seed candidates

3. Pick seeds from candidates

4. Propagate from seeds

CANDIDATE SELECTION METHODS
- Random
- High outdegree
- High inverse PageRank
- High PageRank

Web graph

TrustRank scores

Ranking
Roadmap

- Web spam
  - Definition, techniques
  - Observation
  - Objective

- TrustRank
  - Idea, related work
  - Score propagation
  - Algorithm

- Experiments
Experiments / Data

- **Web data**
  - Entire AltaVista index (June 2003)
  - Site-level web graph
    - 31M nodes
    - 13M without inlinks

- **Seed set**
  - 2500 candidates
  - 178 selected high-quality sites

- **Evaluation sample**
  - 1000 manually tagged sites
Experiments / Sampling

Size

Samples

PageRank

TrustRank

Bucket

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Experiments / Sample Composition

- non-existent 9.6%
- empty 5.6%
- unknown 4.3%
- alias 3.5%
- personal page host 2.2%
- spam 13.5%
- good 61.3%

unusable

usable
Experiments / Spam per Bucket

% Spam

Bucket

PageRank

TrustRank
Experiments / Spam Demotion

Spam from PageRank bucket 3 moved to TrustRank bucket 7
Conclusions

- **TrustRank**
  - Selects good seed pages
  - Propagates scores to other good pages
  - Separates good from bad

- **Experiments**
  - Spam demoted by TrustRank

- **Future work**
  - Different seed selection and score propagation methods

- **Contact:** zoltan@cs.stanford.edu
Inverse PageRank

PageRank $\sim$ inlinks

inverse PageRank $\sim$ outlinks
Experiments / Pairwise Orderedness

Pairwise Orderedness vs. Top–PageRank Sample Sites

TrustRank

PageRank

0.875
0.9
0.925
0.95
0.975

100
200
300
400
500
600
700748

Top–PageRank Sample Sites
Experiments / Precision and Recall

![Graph showing the relationship between TrustRank Bucket and Precision/Recall](image)

**Precision**

**Recall**

The graph illustrates the experimental results of Precision and Recall as a function of TrustRank Bucket. The data points are plotted on a graph with TrustRank Bucket on the x-axis and Precision/Recall on the y-axis. The precision decreases as the TrustRank Bucket increases, while recall increases linearly with TrustRank Bucket.