Web Spam Taxonomy

Zoltán Gyöngyi, Hector Garcia-Molina
Stanford Digital Library Technologies Project, 2004

presented by

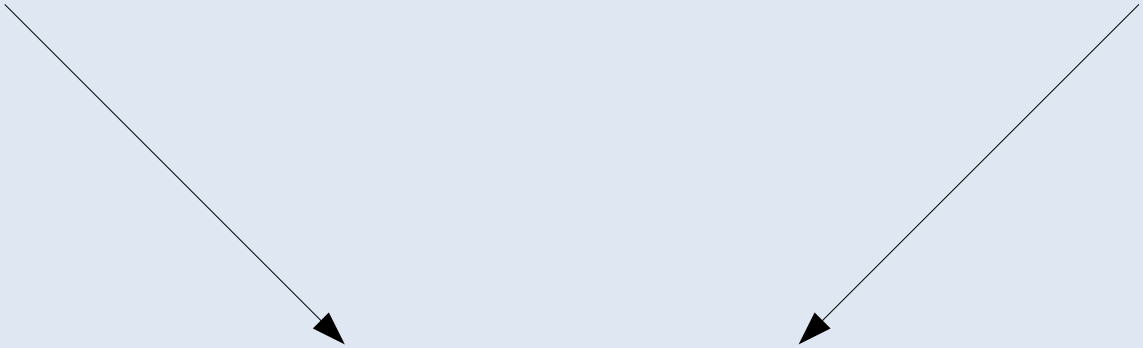
Lorenzo Marcon

Objective of a search engine

Providing results for a query:

relevance
similarity between documents and query

importance
query-independent popularity (e.g. link structure)

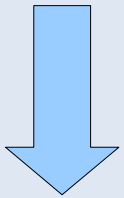


RANKING

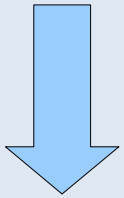
Web spam

Ranking: estimates the *VALUE* of a page

altering the *relevance* or *importance* of a page
without improving its true value



misleading search engine into ranking



SPAM

Boosting techniques and objectives

Term spamming: altering relevance

Link spamming: altering importance

Term spamming (relevance)

Based on TFIDF metric:

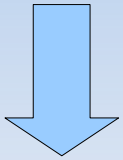
TF(t): *term frequency*, the number of occurrences of term t in the document

IDF(t): *inverse document frequency*, related to the number of document in the collection that contain t

$$\mathrm{TFIDF}(p,q) = \sum_{t \in p \text{ and } t \in q} \mathrm{TF}(t) \cdot \mathrm{IDF}(t)$$

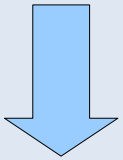
Altering TFIDF

repeating some targeted words



making a page very relevant for a specific query

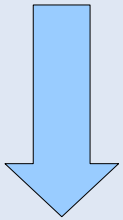
adding a **large number** of distinct terms



making a page relevant for a large number of query

Altering TFIDF

- Spammers don't have real control over terms IDF
- Some search engines ignore IDF scores altogether



increasing TFIDF = increasing TF

Term spamming: where ?

- **Body spam**

  spam terms included in document body

- **Title spam**

  higher weight than body

- **Meta tag spam**

  heavy spamming, low priority

```
<meta name="keywords" content="buy,cheap,
cameras,lens,accessories,nikon,canon">
```

Term spamming: where ?

- **Anchor text spam**

  terms are added to the *target* page

  ```
  <a href="target.html">free, great deals,
  cheap, inexpensive, cheap, free</a>
  ```

- **URL spam**

  url splitted to get page relevance

  ```
  buy-canon-rebel-20d-lens-case.camerasx.com
  buy-nikon-d100-d70-lens-case.camerasx.com
  ```

Term spamming: how ?

- **Repetition**

  terms are repeated to get increased relevance

- **Dumping**

  including large sets of unrelated (and rare) words

easy to filter

Term spamming: how ?

- **Weaving**

  adding spam terms at random position within text

- **Phrase stitching**

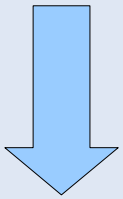
  merging phrases from different contexts or sources

harder to filter!

Link spamming (importance)

Based on web graph structure

Basic idea: people **link** pages they consider **important** on their sites



the value of a page is (also) relative to incoming links

Ranking algorithms: HITS

Assigns two values to each page:

- *Hub score*: links to important authority pages
- *Authority score*: linked by important hubs

circular definition

Ranking algorithms: HITS

increasing ***hub*** score:

Adding many links to important sites (www.cnn.com, www.mit.edu) to the **target page t**.

increasing ***authority*** score:

Increasing the value of *n* hub pages, also adding links to the **target page t**.

Ranking algorithms: PageRank

Ranking factors for a group of Γ pages:

$$PR(\Gamma) = PR_{static}(\Gamma) + PR_{in}(\Gamma) - PR_{out}(\Gamma) - PR_{sink}(\Gamma)$$

$PR_{static}(\Gamma)$: random jump

$PR_{in}(\Gamma)$: incoming links

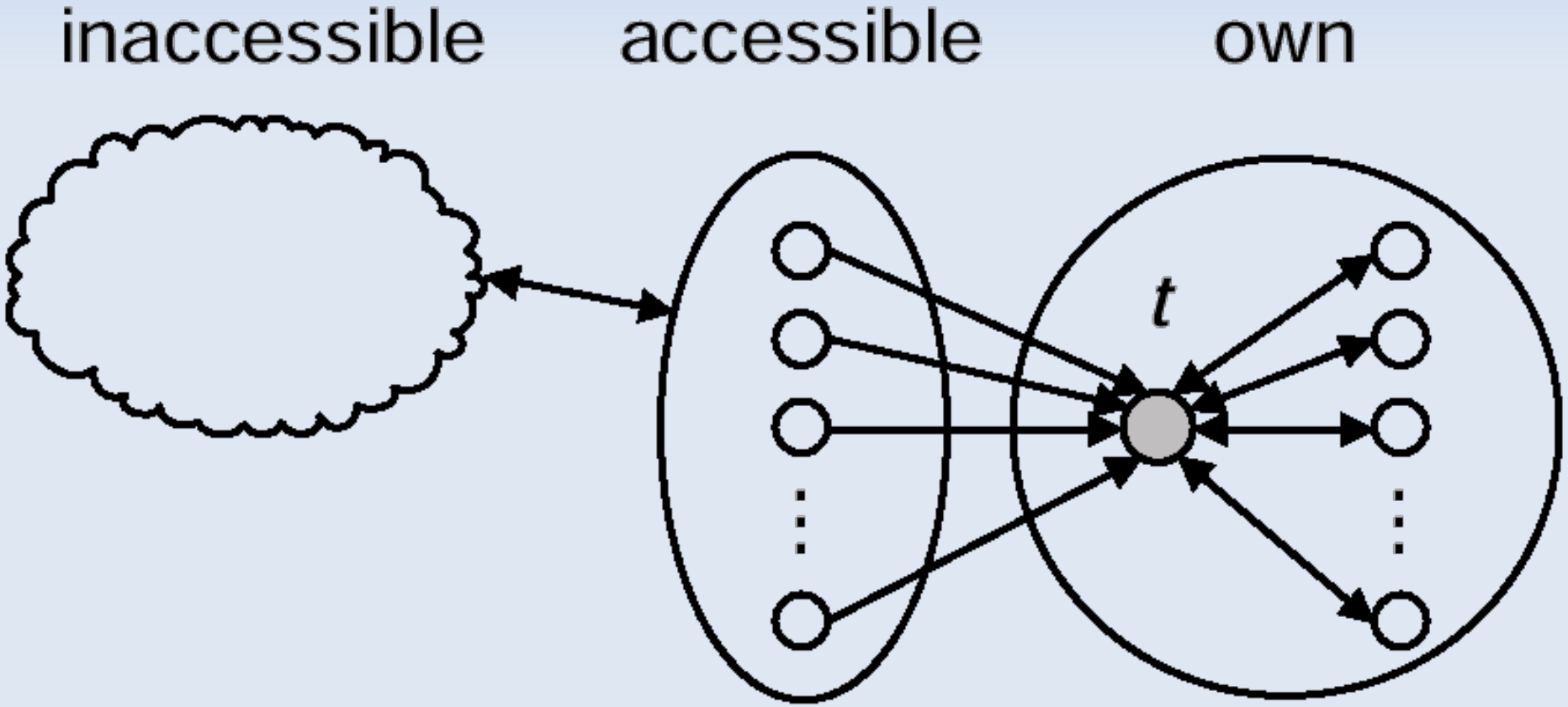
$PR_{out}(\Gamma)$: outgoing links

$PR_{sink}(\Gamma)$: pages without outgoing links

Ranking algorithms: PageRank

Popularity based on number of incoming links

Web from a spammer's eyes:



Link spamming: techniques

Outgoing links:

- Adding links to well known pages
- Directory cloning

Link spamming: techniques

Incoming links:

- Creating a “honey pot”
- Infiltrating a web directory
- Posting links on blogs, guestbook, wikis, etc.
- Link exchanging
- Buying expired domains
- Creating own spam farm

Hiding techniques: content hiding

Examples:

```
<body background="white">
  <font color="white">hidden text</font>
  ...
</body>
```

```
<a href="target.html"><img src="tinyimg.gif"></a>
```

(tinyimg.gif is a 1x1 pixel/transparent image)

```
<script language="JavaScript">
document.getElementById("inv").style.display = none;
</script>
```

Hiding techniques: cloaking

Return a page to regular web browsers, another one to web crawlers

- Mantaining a list of IP used by crawlers
- Filtering user agent

Hiding techniques: redirection

Redirecting URL as soon as the page is loaded

```
<meta http-equiv="refresh"
content="0;url=target.html">
```

easy to parse by SE

```
<script language="javascript">
   location.replace("target.html")
</script>
```

crawlers don't execute scripts

Spam prevalence: statistics

Data set (DS1)

- large set of URLs;
- crawling guided by hash functions;
- manual spam evaluation.

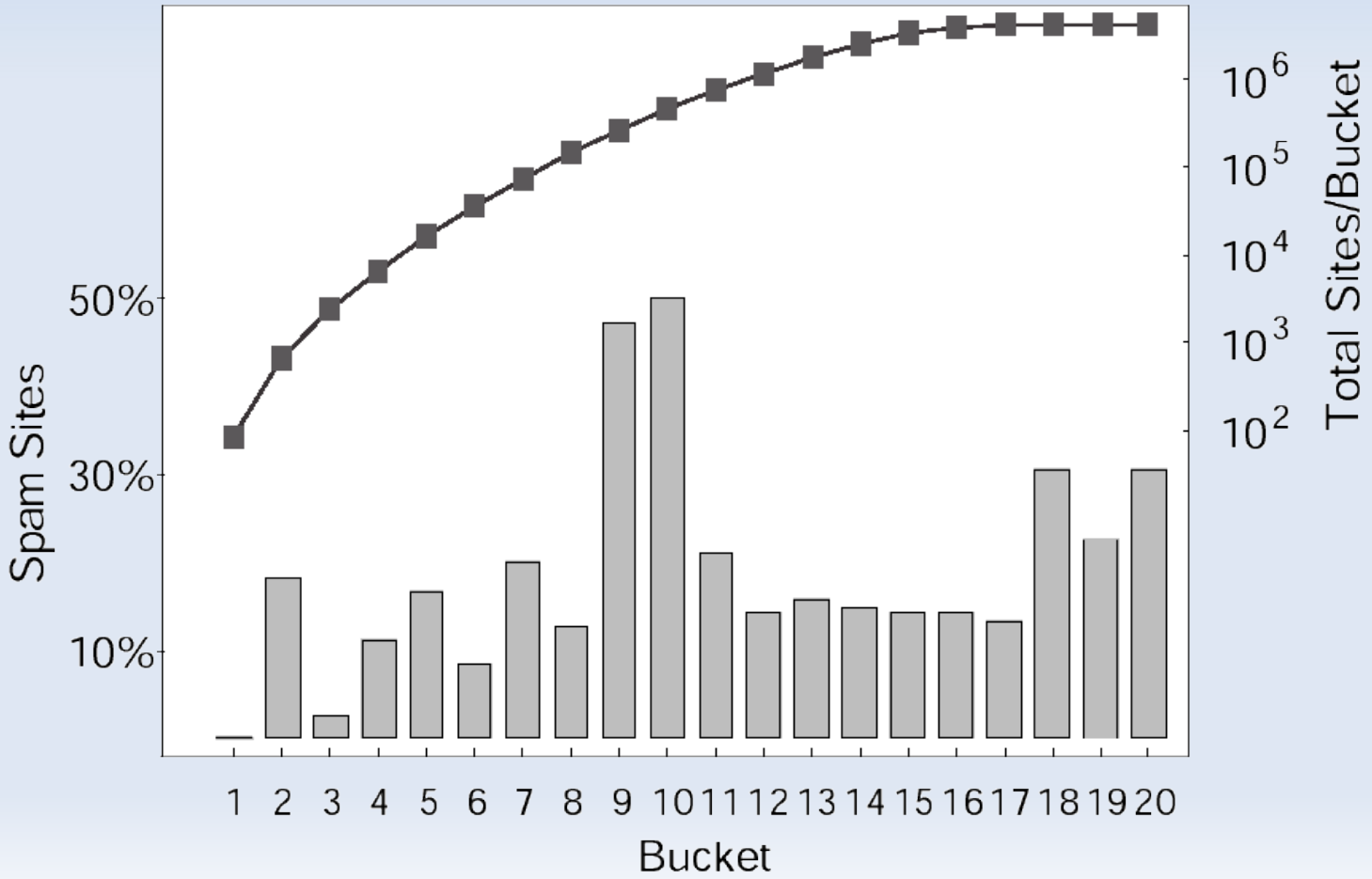
Data set (DS2)

- Single breadth-first search started at the Yahoo! home page;
- manual spam evaluation.

Spam prevalence: statistics

Data set (DS3) – Source: AltaVista.
Pages grouped into about 31 million websites

Sites ordered using descending PageRank, then splitted this list in 20 buckets



Spam prevalence: statistics

Data set	Crawl date	Data set size	Sample size	Spam
DS1	11/02 – 02/03	150 million pages	751 pages	8.1% of pages
DS2	07/02 – 09/02	429 million pages	535 pages	6.9% of pages
DS3	08/03	31 million sites	748 sites	18% of sites

- Crawls performed at different times;
- Different crawling strategies;
- Maybe the average number of pages per site is different for spam and non-spam sites;
- Classification of spam could be subjective.

Conclusions

Web content (estimated): **10-15% SPAM**

Countermeasures:

- *Identification,*
  followed by removal of spam pages from indexes;
- *Prevention,*
  application of various techniques during crawling;
- *Counterbalancing,*
  variation of ranking methods.