Web Spam Taxonomy

Zoltán Gyöngyi, Hector Garcia-Molina

Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005), 2005, Japan.

Apresentado por:
Itamar Hata

Definição

- Ações realizadas para confundir máquinas de busca e levá-las a dar mais prioridade a uma página do que ela merece.
- Também conhecido como:
  Spamdexing, Advesarial Information Retrieval

- Artigo mais atual:
  - Nikita Spirin , Jiawei Han, **Survey on web spam detection: principles and algorithms,** ACM SIGKDD, 2011

Problemas

- Perdas financeiras:
  - $50 bilhões em 2005
  - $130 bilhões em 2009
- Perda de credibilidade da máquina de busca
- Ações maliciosas:
  - Conteúdo adulto
  - Fishing
- # páginas que são spam variam de 6 a 22%.
- Porn sneaks way back on web (1996)

Técnicas Utilizadas pelos Spammers

- Conteúdo
- Elos / URLs
- Consultas Automatizadas

Conteúdo

- TFIDF: apenas a frequência pode ser mofificada. Vetorial, BM25.
  - Repetir as mesmas palavras
  - Acrescentar várias palavras com alto IDF
  - Weaving: acrescentar palavras em conteúdos copiados.
- Fontes:
  - Corpo, Título, meta tag “keywords”, texto ânchora, URL

Google Bomb

- More evil than satan himself → Microsoft (1999)
- Dumb motherfucker → George W. Bush (2000)
- Liar → Tony Blair, (2005)
- Miserable Failure → George W. Bush (2006)
- Plagiator → Victor Ponta, primeiro ministro da Romênia, PhD (2012)

Em 2007 a Google anunciou que modificou seus algoritmos para remover a maioria dos Google Bombs

Elos / URLs

- PageRank, Sergey Brin and Lawrence Page, 1998
  $R(T) = R_{static}(T) + R_{in}(T) – R_{out}(T) – R_{sink}(T)$

- HITS, Jon Kleinberg,1999
  - Hub aponta para várias páginas autoridade.
  - Autoridade: apontada por vários Hubs

Táticas

- Apontar para várias páginas.
- Criar um “honey pot”. Conjunto de páginas que provê algum conteúdo útil.
- Adicionar páginas em diretórios web: yahoo, moz.
- Postar elos em blogs, wikis.
- Hijacking: invadir uma página.
- Trocar elos com outros spammers
- Comprar domínios antigos
- Criar fazendas de spam.

dmoz.org

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

Kids and Teens
Arts, School Time, Teen Life...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

Science
Biology, Psychology, Physics...

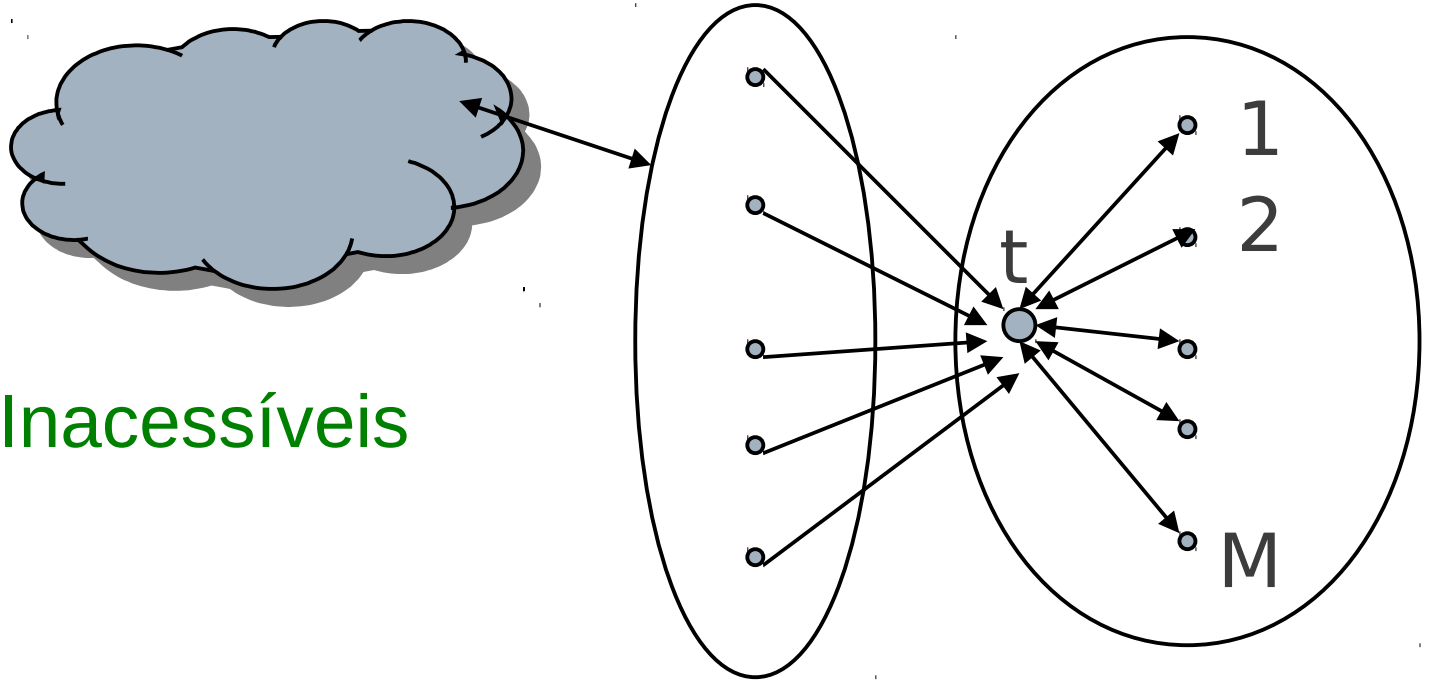
Shopping
Clothing, Food, Gifts...

Society
People, Religion, Issues...

Sports
Baseball, Soccer, Basketball...

World
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

Fazendas de Elos



Inacessíveis

Acessíveis

Fazenda

Outras Técnicas

- Conteúdo escondido: cor da fonte, tamanho, rótulo visible.
- Cloaking: mudança de conteúdo, crawler, user-agent
- Encaminhamento: mudança de conteúdo

Consultas Automatizadas

- Impacta algoritmos baseados em relevance feedback: SVMRank.

- WebPosition: software capaz de realizar esse tipo de consulta.

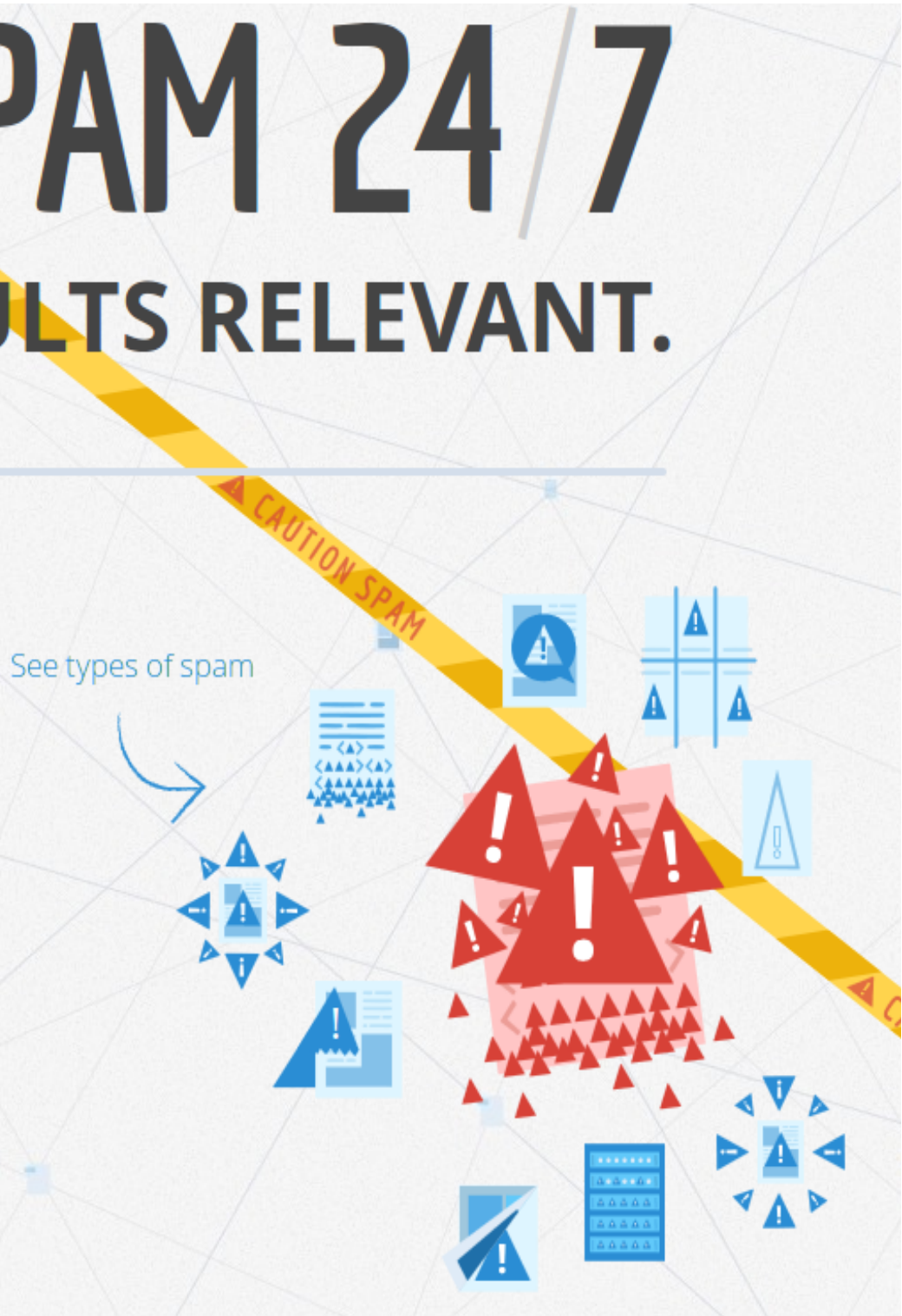
WE FIGHT SPAM 24/7

TO KEEP YOUR RESULTS RELEVANT.

The majority of spam removal is **automatic.** We examine other questionable documents by hand. If we find spam, we take **manual** ACTION.

See what we've removed lately

See action taken over time



Situação Atual - Google

- Conseguem detectar a grande maioria.
- Política de Qualidade
- O restante são análisadas manualmente.
  ~300K páginas são retiradas por mês, 2012.
- Mudanças nos algoritmos de ranking (Panda)
  12% das consultas afetadas, 2012.
- Envio de mensagens aos proprietários dos sites.
  ~400k

Contra Medidas

- Conteúdo, aprendizado de máquina.
  - Distribuição do número de palavras
  - Distribuição do número de elos
  - Conteúdo duplicado
  - Atualização do conteúdo (97%)
- Elos
  - TrustRank, Gyongyi, Garcia-Molina and Pedersen, VLDB, 2004.
    Seleciona páginas confiáveis
    Semelhante ao PageRank.

Conclusões

- Máquinas de busca Web devem considerar Spam.
- 6 a 22% das páginas são spam.
  Existem mais de 30 trilhões de páginas!
- Nem todas páginas podem ser identificadas automaticamente. ~300k são retiradas manualmente no Google por mês.
- Web Spam impactam financeiramente ($130 bi em 2009) e politicamente.