Nullification test collections for Web spam and SEO

Timothy Jones (ANU)
David Hawking (Funnelback)
Ramesh Sankaranarayana (ANU)
Nick Craswel (Microsoft Research)

Detection

Detecting problem content

Nullification

Preventing problem content from negatively affecting search results

The UK-2006 and UK-2007 collections

- Limited to only pages from the .uk TLD
- 80M pages (UK-2006) 100M pages (UK-2007)
- 10k hosts (UK-2006) 100k hosts (UK-2007)
- Labelled partially at the host level
    - Spam
    - Non-spam
    - Borderline
    - Cannot classify

Evaluating nullification with labels

Can test a new ranking by using spam/non spam labels
"Are spam pages demoted by this nullification?"

However, not spam is not the same as relevant

Need to check that relevant pages are not also demoted

Evaluating nullification with users

Preselected information need

- Need to have good answers in the collection
- Collection needs to be relevant to users

User picks information need

- Collection must be current
- Collection needs to be relevant to users

The UK-2006 and UK-2007 collections

Well support testing spam detection

Because of the domain limitation, collection only relevant to UK users

Additionally, the structure may not be representative

- Companion sites
- Popular queries
- Graph statistics

Companion sites in UK-2007

There are many companion sites missing

```
1click-insurance.co.uk -> 1click-insurance.com
1click2keys.co.uk -> 1click2keys.com
1click2keys-overseas.co.uk -> 1click2keys-overseas.com
1click2keysoverseas.co.uk -> 1click2keysoverseas.com
...
3com.co.uk -> 3com.ch,3com.com,3com.cz,3com.de,
   3com.fr, 3com.nl,3com.se
...
abbott.co.uk -> abbott.com,abbott.de,abbott.dk,
   abbott.es, abbott.gr,abbott.ie,abbott.it,
   abbott.no
```

There are 68,000 examples of this in UK-2007

Companion sites in UK-2007

We used a structural heuristic to detect Single Entity Controlled Domains (SECDs) eg news.bbc.co.uk and www.bbc.co.uk

More than 2.4 million non-.uk SECDs referenced

Very high link counts from some .uk domains to non-.uk

11,000 non .uk domains linked to over 100,000 times from a single SECD each

Popular queries in UK-2007

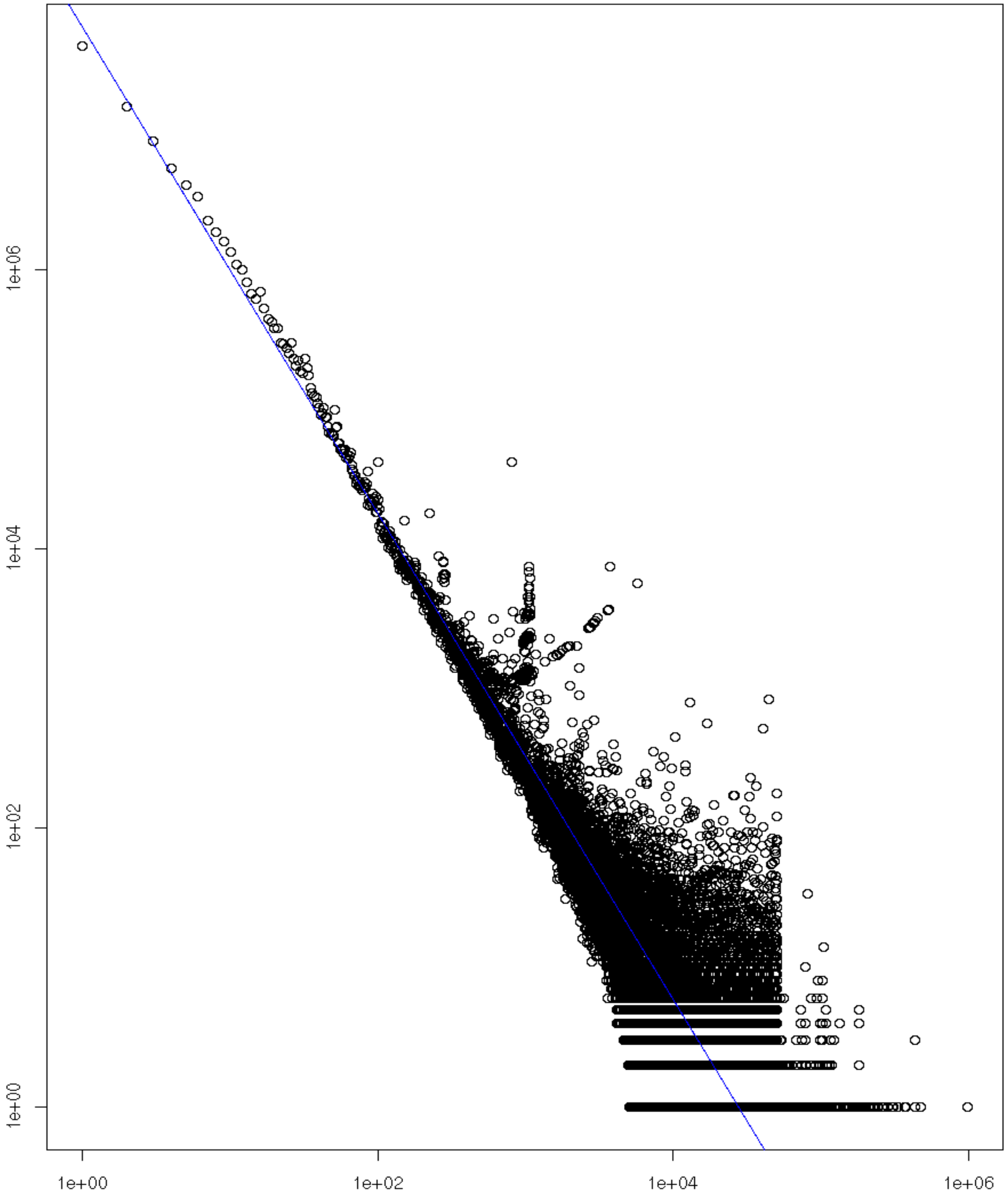
Popular queries are highly targeted by spammers
Submitted the top 10 queries from the UK
(using Google Zeitgeist's year end 2008 list)
To two well known search engines

Only 27.5% of the results were from .uk URLs

Only 17% of the results were present in the collection

Graph structure

Frequency of pages



In-degree

Power law exponent 1.7 when expecting 2.1

The ideal collection

- Page content
- Link graph
- Query and click data

- Large
- Recent

- Spam/non spam labels
- Sample queries known to be targetted by spam
  - With partial relevance judgements or
  - Information need statements for users

Stanford WebBase

Monthly crawls of 61 to 81 million pages

Most recent crawl has 36,000 hosts
35% of the popular query results are present

No spam labels

May not contain much spam

web09-bst

Also known as ClueWeb09, to be used at TREC this year

Contains 1 billion pages

Intentionally designed to contain multi-lingual content

A 50 million page English subset is available

No spam labels

Queries and click data are likely to be available for the TREC web track

http://boston.lti.cs.cmu.edu/Data/web09-bst/

Summary

Evaluation of spam nullification is important, and can't be done with only spam/non-spam labels

Need queries and relevance judgements

Domain limited collections are unlikely to represent the web

For evaluation of nullification, queries, relevance judgements, and a representative web crawl are essential

The web09-bst / ClueWeb09 collection is likely to provide these