Feeding WebSpam Manual Actions to News Corpus

Status: In-progress | Self Link: go/news-webspam | Tracking: b/65543399 | Authors: fatiho@ & aratana@ | Last Updated: 2017/09/15

Background

This document contains the design details for the integration between WebSpam manual actions and G-News corpus. It is agreed that certain manual actions will trigger the removal of the related domain out of News corpus to prevent them being served into downstream news integrations.

Team

Trust & Safety Search	**G-Tech News**	**News Eng**
• Fatih Ozkosemen (fatiho@)	• Adam Ratana (aratana@) • Julie Abbas (julieabbas@)	• Yuqiang Guan (yguan@) • Furkan Tufekci (furkant@)

Option for utilizing webspam badurl files

As per the input from Chris (@chnelson), we investigated the option of utilizing already existing webspam badurls files. Seems like they are not usable for this particular project. Here is our rationale:
References:

- Badurls_demoteindex: List of patterns penalized with demotion penalties
- Badurls_spamindex: List of patterns penalized with removal penalties
- Proto: Shows the scope of these two files as per the different webspam manual actions

Using these two files is risky because:

1. Badurls_spamindex contains patterns for multiple penalties including MobileAdRedirect, CloakingRedirect, SpammyUserContent and HackedIsolated. This makes things complicated; because, corresponding action on the News corpus (complete removal) is harsh and we need to make sure that it is only applied to clear/intentional spam cases. For this reason we only scoped it to Blackhat and Demotion (content+linking) penalties

where we are 100% sure about the webmaster intent. If we use this file, even though it is uncommon, theoretically we may drop a site from corpus that was penalized with SpammyUserContent in the root level. This is something we do not want as the site is innocent in this case. Also we do not want to take action for MobileAdRedirect (sneaky redirects caused by ad scripts) and CloakingRedirect for the time being (maybe at a later stage).

2. Secondly, News Eng will take a different action for hacked cases: Temporary removal from corpus until the hack is cleaned. Having the HackedIsolated penalties in Badurls_spamindex that also contains Blackhat cases makes things complicated as there is no information to differentiate them. This file only contain url pattern info.

We are, instead, moving forward with a custom design that exports desired penalties in csv to be consumed by News Eng.

Scope

We are moving forward with creating three custom .csv files for integration as this is the most convenient way for the Eng team to consume. Here are the details:

- **Data Source for Webspam Manual Actions:** squeal.patterns
- **Penalties in Scope:** Only removal, demotion and hacked penalties are in scope below is the full list and some related details

	Removal	**Demotion**	**Hack**
Penalty Type	Blackhat	DemoteForContent DemoteForBackLinks DemoteForForwardLinks (ex. Bad content - scraping from multiple sources, no sign of human curation - **100%** sure of bad webmaster intention)	HackedURLs (injected at root - demotes for spammy queries “viagra” - shows warning to user “site may be attacked” / chrome interstitial) -- this IS served in Google News, no warning HackedIsolated (injected urls - removed from search index - also disappear from Google News) HackedHidden (injected content into organic existing URLs - end user doesn’t see the content - only visible to googlebot / cloaking) HackedRedirect (happens in Google

			News - user clicks URL, but gets redirected to hacker's own site, ie canadian-pharma.com)
Example Site	screenshot	screenshot	screenshot
Corpus Action	Complete Removal (Rejected)		Temp. Removal
Output	webspam_root_blackhat.csv, webspam_root_demote.csv, webspam_badurl_patterns.csv		webspam_badurl_patterns.csv
Frequency	Daily		Daily

- **Partial Manual Actions:** Such as subdomain1.example.com or example.com/subfolder1/ or a deep url like example.com/nicepage.html
    - Majority are ascribed to root domain, when possible/isolatable, can sometimes ascribe more granular Ex: fr.example.com vs example.com
    - We will annotate the data with a flag (is_root_penalty) and let the News Eng decide on the action
- **Pattern structure:** Open (example.com) or closed (example.com/)?
    - Go ahead with the standard penalization pattern.
- **Sitechunking:** How to cover free hosts (sitechunks)? E.g. examplenewssite.wordpress.com/subfolder/news.html
    - We will use webspam_sitechunk_utils for sitechunking. go/plx-sitechunk
    - This will give us the correct sitename as an anchor to join news corpus and squeal.patterns
- **Frequency:**
    - TSS exports twice a day
    - Daily output is OK for us
    - Borgcron / Workflow schedule to run every 8 hours to ensure fresh content.
- **Joining News Corpus and Manual Actions:** Best way to join squeal.patterns (updated almost instantly) and news DB
    - We (gTech) will setup a Borgron to:
        - Convert gnews_newsish_all.sstable to capacitor
        - Run dremel query against squeal.patterns
        - Output csv file that news-eng can read from
    - SoD access for mdb.dpub-bi-news has been granted
    - CL in progress: cl/168277655 | Related bug: b/65543399

- **Output Structure**
    - Pattern (e.g. subdomain.example.wordpress.com/subfolder/)
        - A subfolder / pattern may belong to a separate publisher, so this represents the most granular level of penalty

        - Patterns are normalized to:
            - regexp:^http://www\\.autorevue\\.cz/.*\\?forum
    - Site Name (eg subdomain.example.com/)
        - Separate sites can exist as subdomains
    - Domain (e.g. example.com/)
        - Closed pattern (trailing / to avoid regex matching with say, .computer)

Outputs

- Access via mdb/dpub-bi-news-readers
- Dashboard visualization: go/news-corpus-webspam
- CSV: **Sources for Rejection**:
    - Contains sites with Blackhat + Demote root penalties
    - Columns (with header row):
        - source_id
        - source_url
        - normalized_penalty_pattern
        - is_root_penalty
        - is_golden_supergolden
        - Is_newsish
- CSV: **BadURL Patterns (don't serve)**:
    - 1 column (no header row):
        - normalized_penalty_pattern
    - Contains root penalties for hacked sites OR non-root penalties, but not both. Root penalties cover non-root, so no need to include non-root.
    - Contains root Blackhat + Demote penalties

Sources to Reject	**Location** (has header row)
Blackhat	`/placer/prod/home/dpub-bi-news/gnews/webspam/reject/blackhat/webspam_root_blackhat-00000-of-00001.csv`
Demotion	`/placer/prod/home/dpub-bi-news/gnews/webspam/reject/demote/webspam_root_demote-00000-of-00001.csv`

BadURL Patterns	**Location** (no header row)
Hack, Blackhat, Demotion	`/placer/prod/home/dpub-bi-news/gnews/webspam/badurl/webspam_badurl_patterns-00000-of-00001.csv`