

Web spam

Adam Kilgarriff

Lexical Computing Ltd. &
University of Leeds
UK

adam@lexmasterclass.com

Vít Suchomel

Lexical Computing Ltd. &
Masaryk University
Brno, Czech Republic

vit.suchomel@sketchengine.co.uk

Abstract

Web spam is getting worse. The biggest difference between our 2008 and 2012 corpora, both crawled in the same way, is web spam. In this paper we talk about what it is, with examples and a discussion of the overlap with ‘legitimate’ marketing material, and present some ideas about how we might identify it automatically in order to filter it out of our web corpora.

1 Introduction

Web spamming “refers to actions intended to mislead search engines into ranking some pages higher than they deserve” (Gyöngyi and Garcia-Molina, 2005). Web spam is a problem for web corpus builders because it is quite like the material we want to gather, but we do not want it. (We assume a ‘general crawling’ method for web corpus construction.)

Here are some examples:

The particular Moroccan oil could very well moisturize dry skin handing it out an even make-up including easier different textures.

Now on the web stores are very aggressive price smart so there genuinely isn’t any very good cause to go way out of your way to get the presents (unless of course of program you procrastinated).

Hemorrhoids sickliness is incorrect to be considered as a lethiferous malaise even though shut-ins are struck with calamitous tantrums of agonizing hazards, bulging soreness and irritating psoriasis.

It is on the increase: when we compare two corpora gathered using the same methods in 2008 and 2012, enTenTen08 and enTenTen12, the web spam in the later one is the most striking difference.

It is a moving target. The spammers and the search engines are in a game where the spammers invent new techniques, which will often work for a while until the search engines have worked out

how to block them. Meanwhile the spammers will work out new techniques. The comments in this paper are likely to be of purely historical interest in the near future.

Our concern for web spam has been driven by specific corpus studies (all for English). In one we were investigating the term “Moroccan oil”. In enTenTen08 it scarcely occurred, in enTenTen12 most occurrences were spam associated with the beauty products industry. In another we were investigating “on the ? store” and found that most instances for four of the top fillers for the variable slot, *web*, *net*, *internet*, *online*, were spam. In a third we were looking into rare words found in dictionaries, and checked in enTenTen12 for a word we did not know, *lethiferous*. Twelve of its fourteen instances in enTenTen12 were spam.

Most web corpus builders use a range of filtering strategies such as checking that documents have mostly common words, and a plausible proportion of grammar words: web spam that was not fairly similar to good text would largely be filtered out by these processes. The remaining web spam looks quite like good text.

1.1 Intermediate cases

Consider the text chunk below:

MoroccanOil is an oil treatment for all hair types. Moroccan Oil is alcohol-free and has a patented weightless formula with no build up. Softens thick unmanageable hair and restores shine and softness to dull lifeless hair. Instantly absorbed into the hair. Moroccan Oil will help eliminate frizz, speeds up styling time by 40%, and provides long-term conditioning to all hair types. Are \$20 shampoos and conditioners worth it? Can good hair-care products be found at the drugstore, or are the expensive salon products really superior? In this comprehensive guide to all things hair care,

Taken on its own this is respectable English. However there were many such pages, often with the same short sentences and sentence fragments in

different order or mixed in with less coherent and grammatical parts, often also on pages of “news items” with a ‘read more’ link at the end of each paragraph. The text is a marketing text, with component sentences written by a person, but that does not exclude it from being spam (on the definition we opened with). The line between marketing and spam is not easy to draw.

A recent development in this territory is ‘content farms’ where people are paid (poorly) for writing lots of articles, with the primary goal of driving traffic to advertising sites.¹ This is human-written and coherent, yet fits our definition of web spam. It is not clear whether we want it in a linguistic corpus.

2 Related work

(Gyöngyi and Garcia-Molina, 2005) present a useful taxonomy of web spam, and corresponding strategies used to make it. Their paper was presented at the first AIRWeb (Adversarial Information Retrieval on the Web) workshop: it was the first of five annual workshops, associated with two shared tasks or ‘Web Spam Challenges’. The last of the AIRWeb workshops was 2009; in the years since, there have been joint WICOW/AIRWeb Workshops on Web Quality.² These workshops, held at WWW conferences, have been the main venue for IR work on web spam.

Since the merge, there has been less work on web spam, with the focus, insofar as it relates to spam, moving to spam in social networks and tagging systems (Erdélyi et al., 2012).

The datasets used for the shared tasks are called WEBSPAM-UK2006 and 2007 and are described in (Castillo et al., 2008). Labels (spam or non-spam) were at the level of the host rather than the web page. A large number of hosts were labelled in a substantial, collective labelling effort: 7473 hosts in UK2006 and 6,479 in UK2007. UK2006 had 26% spam whereas UK2007 had 6% spam: the difference is because UK2006 did not use uniform random sampling of a crawl whereas UK2007 did, so 6% is the useful figure for reference. The tagged data was split with two thirds usable for training, one third retained for evaluation. There were six participants for UK2007 and all used supervised machine learning, with a range

of text-based and link-based features, and the best system scoring 85% ‘area under curve’. This was improved upon by (Erdélyi et al., 2012), who also discuss the ECML/PKDD Discovery Challenge dataset where ‘spam’ is one of a number of labels.

2.1 Search Engines

Web spam is a game played between spammers and search engines. Search engines—particularly the market leader Google, also Bing, Yandex, Baidu—employ teams of analysts and programmers to combat spam. In those companies there will be great knowledge of it and expertise in identifying it. They probably have large recent databases of spam, to conduct experiments on. However these resources and expertise will not, for obvious reasons, be shared outside the company. A good feature of AIRWeb is that representation on it from search engine companies is high: Carlos Castillo, from Yahoo, notes in his powerpoint reviewing the Web Spam Challenges³ “keeping web data flowing into universities” as a goal and a benefit of the Web Spam Challenge.

The Google paper “Fighting Spam”⁴ describes in broad terms the kinds of spam that Google finds, and what they do about it. Figure 1 shows developments from 2004 to 2012.

The BootCaT method for building corpora (Baroni and Bernardini, 2004) works by sending seed terms to a search engine, and gathering the pages found by the search engine. In this approach, the corpus-builder benefits directly from the search engine’s measures against web spam.

2.2 Test data and evaluation

It is a big methodological challenge to gather a good sample of web spam. It is, by design, hard to find and set apart from good text. We can gather samples by simply noticing and putting spam documents to one side to build up a spam corpus. This is useful and probably central to all we might do, however it does not help us find the spam types we have not yet noticed.

Historical datasets are of limited value as spammers will have moved on: despite that, the WAC community will almost certainly benefit from using the AIRWeb and ECML/PKDD datasets discussed above, and the filtering methods developed

¹http://readwrite.com/2010/11/17/content_farms_top_trends_of_2010

²WICOW stands for “Workshop on Information Credibility on the Web”.

³<http://airweb.cse.lehigh.edu/2009/slides/castillo-challenges.pdf>

⁴<http://www.google.com/insidesearch/howsearchworks/fighting-spam.html>

Manual Action by Month

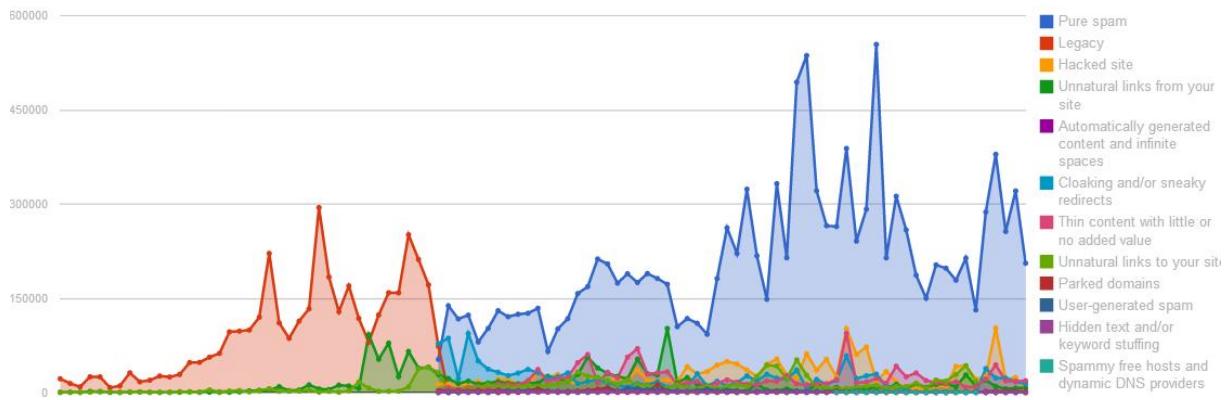


Figure 1: Google’s analysis of spam types and quantities, 2004-2012.

there.

2.3 Level of analysis

The IR work mostly focuses on finding bad hosts (and much of it, on links, “the web as a graph”). That is a distinct strategy to finding bad text, e.g. within a web corpus once it has been cleaned, with links deleted. One question for web corpus builders is: at what stage should spam detection take place - before html-removal, or after, and do we work at the level of the page or the website? Also, should we concentrate on hosts, or domains, or web pages? Some preliminary evidence suggests that the landscape hosts and domains change very quickly, so methods based on text may retain validity for longer.

3 Methods

3.1 Coherence approaches

To recognise the examples above as web spam, we have to read them. This is in contrast to, for instance, noticing unwanted material not in English, or lists of English words, where a cursory glance is sufficient and the level of attention that deserves the word ‘reading’ is not required. The spam is not obviously grammatically flawed. But it lacks coherence. This suggests that, to identify it, we want to measure the coherence of each sentence or text, in order to identify spam as the low-scoring material.

Ways in which we might do this are:

- apply the entity-grid model of (Barzilay and Lapata, 2008);

- perform syntactic analysis to create dependency trees to model dependencies of parts of sentences. A “nice” tree could mean the sentence is coherent;
- in a coherent text we expect words to be from compatible domains and registers. It may be possible to identify sets of words that belong together (in terms of domain or register) and then to spot texts where words come from mismatched or incompatible domains or registers.

3.2 Words for things that people want to sell, and marketing buzzwords

Much web spam works to sell products, so the names of the things being sold often be mentioned, as in Moroccan oil. Spammers will also use low-content terms that they think people will search such as “web store”. If we can gather a long list of these items, we can use counts for them as part of a scoring system. (Baisa and Suchomel, 2012) explore this method, using a small spam corpus to identify n-grams which are notably more frequent there than in reference text.

3.3 Dictionary words

Lethiferous points to spammers using dictionaries to flesh out the linguistic profile of their spam. Perhaps texts containing words which are in big, traditional dictionaries but have low corpus frequencies can act as alarm bells.

For removal of duplicates and near-duplicates in our corpora we use onion (Pomíkálek, 2011). However we have recently noted that some web

spam avoids detection through random changing of content words to synonyms, drawn from a thesaurus. This is a method that could be reverse-engineered.

4 EnTenTen12 vs. EnTenTen08

We stated above that the biggest difference between EnTenTen12 and EnTenTen08 is web spam.

(Kilgarriff, 2012) presents a method for exploring differences between corpora, demonstrating how the manual classification of the top 100 keywords of corpus1 vs. corpus2 and vice versa gives a rich picture of the contrasts between the two. This is what we have done in this case, as follows:

- For each word matching
 - Find frequencies in corpus1 and corpus2
 - Normalise to ‘per million’
 - Add a ‘simplemaths parameter’ of 0.001 to normalised figures (including the zeroes). This low value for the parameter means that the list will be dominated by low-frequency keywords.⁵
 - if the figure for corpus1 is larger than that for corpus2, divide the corpus1 figure by the corpus2 figure to give a score
- sort the words according to the scores

The highest-scoring words are the keywords for corpus1 vs. corpus.

One typically finds many names and nonwords in the lists so generated, and we were interested in dictionary words. We filtered to give only all-lower-case-letters items of length at least 3, and hunspell,⁶ to give a list of words that were only ‘dictionary words’. Table 1 shows the top of the list complete with frequency figures, to show the sheer magnitude of the differences in frequencies: 18,102 occurrences of *jewelries* in 2012 against 35 in 2008. Table 2 gives the full analysis.

Of the 100 words, six related to new things: three (*tweeting tweeted twitter*) to twitter, launched in 2006 with meteoric growth since 2007; *voltaic*, almost always in the context of photo voltaic cells, newly topical with climate change and associated government initiatives; *atomizer*, for which all the data related to electronic cigarettes (of which an atomizer is one part),

which first appeared on the international market in 2005-06,⁷; and *jailbreak* which is what you do when you convert an Apple device such as an iPhone or iPod from one that can only operate in the Apple-approved ways to a general purpose device. In addition there was one new word, *colorway* (in both singular and plural; a synonym, widely used by clothing and footwear manufacturers, for *colour scheme*: “we have this design in all sizes and colorways”) and *aftereffect*, increasingly spelt as one word.

Of these, *atomizer* and *colorway* relate to things that are marketed extensively on the web. So do most of the other 91 items. The straightforward shopping items are clocks and watches (six words), footwear (five), handbags and holdalls, birthstones (singular and plural), *pantyliners*, *jerseys*, *headpins* and *foodstuffs*. Services were financial (six items), locksmiths (two), *refacing* for kitchen cabinets and four words relating to weddings.

‘Health and beauty’ accounted for 28 of the 100 keywords. The leading subcategory is skin, with particular emphasis on spots. We have pimples, blackheads, whiteheads, moisturizers and dehydrators. The meaning of *breakouts* that put it in the keyword list was “a breakout of acne” and a *concealer* was always a concealer of acne.

There were just two items of a *lethiferous* flavour: *accouter* and *osculate*. *Accouter*, a rare synonym for *dress* (as in *accoutrements*) was widely used in spam associated with clothes and weddings. *Osculate*, a rare synonym for *kiss*, in spam associated with pornography.

The remaining large category was formed of words in morphological forms that were unusual for them: eleven nouns ending in ‘-ness’, six plurals, two nouns and an adjective in -er, and two adjectives with -able.

The -ness nouns included *humorousness*, *severeness*, *comfortableness*, *anxiousness*, *courageousness*, *neglectfulness*, *safeness*. These are odd because it is usual to use *humo(u)r*, *severity*, *comfort*, *anxiety*, *courage*, *neglect*, *safety* instead.

The plurals include mass nouns *attire*, *apparel*, *jewelry* which, in the first author’s British dialect, scarcely bear pluralising at all.

The items *anticlimaxes*, *dejecting*, *unexceptionally* all have something contradictory about them. An anticlimax only exists in contrast to an ex-

⁵See (Kilgarriff, 2009) for discussion.

⁶<http://hunspell.sourceforge.net/>

⁷Wikipedia: *Electronic cigarette*

Word	enTenTen12		enTenTen08		Score
	Freq	Norm	Freq	Norm	
tweeted	28711	2.2	11	0.0	507.41
jewelries	18012	1.4	35	0.0	118.72
tweeting	26024	2.0	67	0.0	93.40
colorway	6395	0.5	17	0.0	79.69
hemorrhoid	57951	4.5	181	0.1	79.29
straighteners	28206	2.2	133	0.0	52.20
courageousness	8717	0.7	40	0.0	50.86
twitter	712447	54.9	3602	1.1	49.81
straightener	23324	1.8	137	0.0	41.94
colorways	4242	0.3	23	0.0	40.83
anticlimaxes	2584	0.2	14	0.0	37.91
wagerer	1060	0.1	4	0.0	37.21

Table 1: enTenTen12 top keywords, showing figures and working.

pected climax, and climaxes tend to be singular by their nature, so it is hard to see a role for the plural version of their contrasts. The verb *deject* is always passive so it is hard to see how something can be dejecting. *Exceptionally* brings attention to the predicate it is associated with: when we negate it with -un it is unclear what we are doing.

Discussion

All 100 words except the three twitter words and *voltaic* were highly associated with spam, as confirmed by scanning concordances. For some –*wagerer, osculate, conveyable*– all of a sample of fifty concordance lines appeared to be spam, but for the majority, the judgement was not easily made, with most of the sample being on the spectrum between marketing and gibberish.

For the shopping, services, and health-and-beauty words, we see the results of spammers taking legitimate material, chopping it into pieces and permuting and varying it.

The morphology cases are more puzzling. Three hypotheses for the radical increases in frequency of these terms are:

1. A computer is generating derived forms of words and using them in spam: example

This, in addendum to modern sedate safety concerns, numberless increases in data sum total, and rising cost pressures, closest these organizations with some uncommonly outstanding topic challenges.

2. Authors are non-native speakers of English. They will often use the regular nominalisation (*anxiousness*) rather than the irregular one (*anxiety*) and pluralise mass nouns in er-

ror. The following seems likely to be a non-native production:

The minimum height I would suggest for your inside rabbit cage would be 40 cm, but this only a guide. Please use you discretion and if in doubt go for the taller cage. A lot of individuals choose for numerous floor bunny rabbit cages with brings joining the levels. This grants the bunny rabbit a lot extra room without borrowing more room inside your haven. Owning a line flooring inside your bunny rabbit Cage isn't a good plan if you would like to give **comfortableness** for your bunny rabbit. While having a wire bed with a pull out and makes for simpler maintaining, it's not all of the time necessary as bunnies are easily litter box trained.

3. It is a matter of dialect: whereas the first author will always say *comfort* rather than *comfortableness*, and for him, *jewelries* is close to impossible, this is not so in other dialects. (Kachru, 1990) discusses the varieties of English in terms of the *inner circle* (the traditional bases of English: UK, USA, Australia, New Zealand, Ireland, anglophone Canada), the *outer circle*: countries where English is historically important and is central to the nation's institutions; South Africa, India, Nigeria, the Philippines, Bangladesh, Pakistan, Malaysia, Kenya; and the *expanding circle*, where English is playing a growing role, which covers much of the rest of the world. The inner circle countries are all high-wage, so it would not be surprising if companies looked to outer-circle countries, where there are both many speakers of local dialects

NEW THINGS

tweeting tweeted twitter
(photo) voltaic (cells)
atomizer (as part of apparatus for giving up smoking)
jailbreak (verb: remove limitations on an Apple device)

NEW WORDS

colorway colorways aftereffect (increasingly spelt as one word)

SHOPPING

footwear espadrille sneaker slingback huarache
handbags holdalls
chronograph chronographs timepiece timepieces watchstrap watchmaking
birthstone birthstones
foodstuff
headpins (jewelry making)
pantyliner jerseys

SERVICES

locksmith locksmiths refacing (for kitchen cabinets)

MONEY

refinance refinancing remortgages defrayal cosigner loaners

WEDDINGS

bridesmaid boutonnieres honeymoons groomsmen

HEALTH AND BEAUTY

periodontist whitening veneers aligners (both mainly for teeth)
hemorrhoid hemorrhoids
hairstyles straightener straighteners
slimming physique cellulite liposuction stretchmarks suntanning
moisturize moisturizes moisturized dehydrators detoxing
pimples whiteheads blackhead blackheads
breakouts (of acne etc) concealer concealers (of acne etc)
tinnitus

RARE DICTIONARY WORDS

accouter osculate

MORPHOLOGY

humorousness severeness sturdiness impecuniousness comfortableness
anxiousness adorableness courageousness neglectfulness moldiness safeness
anticlimaxes chitchats attires apparels jewelries jackpots
wagerer vacationer dandier
acquirable conveyable
dejecting unexceptionally

NAMES (incorrectly included - most were filtered out)

spellbinders (company) circuital (album) android (operating system)

OTHER

frontward proficiently

Table 2: An analysis of the top 100 keywords of enTenTen12 vs. enTenTen08 (simplemaths parameter=0.001, filtered to give only all-lowercase dictionary words at least three characters long). All capitalised text is authors' labels for categories, and all text in brackets is explanatory glosses. All other words are the keywords.

of English, and low wages, to write bulk marketing material for SEO. Consider:

It is dream of every woman to have a perfect wardrobe. The thing that tops the list to make the wardrobe a complete one is a black shoe. Ladies black shoes add style and versatility to the **attires**. From casuals to formal black is the colour that makes the feet stand out from the crowd.

To the first author's British ear, this sounds like Indian English.

5 In sum

Web spam is a large and growing problem for web corpus builders, at least for English. There has been work on it in the IR community (to date, to the best of my knowledge, not known to the WAC community). The WAC community can benefit from that work.

We have also presented some linguistic observations that could prove useful for spam identification, and some data relating to changes we have observed between 2008 and 2012.

6 Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013. The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is highly appreciated.

References

- Vít Baisa and Vít Suchomel. 2012. Detecting spam content in web corpora. In *Recent Advances in Slavonic Natural Language Processing (RASLAN-6)*, Masaryk University, Brno, Czech Republic.
- Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, pages 1313–1316.
- Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- C. Castillo, K. Chellapilla, and L. Denoyer. 2008. Web spam challenge 2008. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.
- Miklós Erdélyi, András Grazo, and András A. Benczúr. 2012. Web spam classification: a few features worth more. In *Proc. Joint WICOW/AIRWeb Workshop at WWW-2012*.
- Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Web spam taxonomy. In *AIRWeb, Proceedings of a Workshop on Adversarial Information Retrieval on the Web*, pages 39–47.
- Braj Kachru. 1990. *The alchemy of English: the spread, functions, and models of non-native Englishes*. University of Illinois Press.
- Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Int. Conf. Corpus Linguistics*, Liverpool.
- Adam Kilgarriff. 2012. Getting to know your corpus. In *Proc. Int. Conf. Text, Speech, Dialogue*, Brno, Czech Republic.
- Jan Pomikálek. 2011. *Removing Duplicate and Boilerplate Content from Web Corpora*. Ph.D. thesis, Masaryk University, Brno, Czech Republic.