

Spam Recognition Guide for Raters

for your convenience the changes are highlighted with UPDATE throughout the document

Introduction

During the course of rating, you may encounter results that Google considers spam. Some are obvious but others are less overt. Provided here is an overview of spam recognition tools for use in rating projects.

Before familiarizing yourself with tools aimed at detecting spam, i.e. deceitful web design, please read Google's policies on quality web design <http://www.google.com/webmasters/guidelines.html#quality> . In particular, pay attention to:

- The distinction between pages designed for human viewers and those set up for search engine robots
- The specific enumerated manipulative techniques for which sites may be “punished” by Google.

If you are not sure of your spam detection skills yet, you may want to subject *every* result page that comes up for rating to a checklist of all potential manipulative techniques that this guide explicates. With experience in spam identification, the spam-spotting techniques presented below become easy to use. You will have seen patterns of honest pages and deceitful pages; questionable results will jump at you “asking” to be checked for evidence of spamming. If unsure, do not hesitate to ask questions!

Note on Foreign Language spam: If a page in another language uses an obvious spamming technique, do label it as spam. Spam identification often does not depend on linguistic issues. However, if you are unable to make a determination, feel free to rate the result as Foreign Language. The same logic applies to Offensive pornographic results that are neither invited nor tolerated by the query. If you can make determination independent of the language, please do so.

Common Spam Techniques

Sneaky Redirects

What you'll see on your Quest page: URL A is shown as a query result.

When you click on the link: URL A may appear in the address bar of the browser for a brief moment, but you are sent to URL B. You might see other, transient URLs before the page finally loads with URL

B visible in the address bar. One URL may sneakily redirect to a number of rotating domains, so clicking on the same result several times may land you on pages under different URLs. Those pages may or may not look the same.

What's probably going on: Domain B wants to extend its reach in our index, so it creates Domain A. Google indexes and scores the content on Domain A, yet the user is redirected to Domain B. The webmaster presents one content to the search engine robot and another to the users.

Examples:

Result URL	What visiting the page takes you to ¹
http://www.lasik-eye-surgery-laser-eye-surgery.com/	http://1800contacts.com/ or http://www.visiondirect.com/spanish/scripts/default.asp?AID=9483447&PID=858188
http://www.juvenews.com/pics-of-car-wrecks.html	http://www.ofhg.com/sexsites/index.html

<http://www.theii.net/information-on-the-great-pyramid-at-giza.html>

<http://www.scbgalleries.com/freeporn/index.html>

or <http://www.scbgalleries.com/pornogallery/index.html>,

or <http://www.scbgalleries.com/openadult/index.html>,

or

<http://www.ofhg.com/freadult/index.html>,

<http://www.ofhg.com/gallery/welcome.html>,

<http://www.ofhg.com/sexsites/index.html>...

<http://pregnancy.pregnancy-pampers.co.uk/>

http://uk.pampers.com/en_GB/signup.do

<http://children.pregnancy-pampers.co.uk/>

http://uk.pampers.com/en_GB/signup.do

¹

. Hotlinks have been disabled for some porn pages whose content is apparent from the URL structure.

Question: Are all redirects spam?

Answer: Absolutely not!

For example, <http://www.film.com> redirects to movies.real.com, but not in a sneaky manner.

For another example, consider www.compaq.com. Compaq is now a Hewlett Packard company.

www.compaq.com redirects to <http://h18000.www1.hp.com/> in a legitimate manner.

100% Frame

What you'll see on your Quest page: URL A is shown as a query result.

When you click the link: URL A appears in the address bar of the browser. The page uses a frame that occupies all (or nearly all) of the browser window. Page B fills this frame. You need to reveal the page information for page B. In Internet Explorer, point to any place on the main page (other than an image) inside the frame with your cursor, right-click and choose "Properties". Check Address: (URL).²

What's probably going on: Domain B is a legitimate commercial site that wants to extend its reach in Google's index, so it creates Domain A. Google indexes and scores the content on A, yet the user is shown Domain B in the 100% frame. Again, what's created for search engine robots differs from what is created for human visitors.

Example: <http://www.catwalk4u.de/> (right-click on the web page body and choose "Properties" in IE, and note the URL, which may be one of a number of rotating sites, including <http://www.link-diener.de/mode.html> , <http://www.trixo.de/mode.html> and <http://www.looking4links.de/mode.html>).

Hidden Text / Hidden Links

What you'll see on the result page: You may notice large blank areas on the bottom or/and the top of the page. Using the keyboard shortcut for Select All on the page (CTRL-A in Internet Explorer) may reveal text or links that are hidden from the user (example: white text on white background).

²

Certain pages, primarily those that contain objects that can be copied, disable this feature.

What's probably going on: The webmaster hopes that adding more text to the page will increase the number of ways in which users can find the page searching on Google. Stuffing the page with text may put off site visitors, so the webmaster chooses to hide the text and/or links. Google scores content that the user never sees; what's being created for search engine robots differs from what is intended for human page viewers.

Example 1: <http://www.marantz.com/> -- observe pristine white space and then do select-all to reveal white-on-white text.

Example 2. On the bottom of these pages observe hidden text in a very small font size:

<http://www.jobjobbed.com/>

http://free-web-hosting-inc.com/fort_wayne_indiana_web_hosting.html

Porn on Expired Domains

What you'll see on your Quest page: URL A is shown as a query result. It has a relatively “benign” domain name, with no reference to porn or adult content.

When you click the link: The page has porn content.

What's probably going on: An adult content webmaster purchased Domain A after its former owner allowed his/her ownership to lapse. In Google, Domain A has some lingering good reputation in the form of PageRank. Webmasters linking to Domain A aren't always on top of their links, and their “votes” for Domain A based on old, benign content can continue indefinitely, to the adult content webmaster's benefit. Google is counting incoming hyperlinks that the new, adult content webmaster never earned, and search relevancy can be skewed.

Secondary Search Results / PPC

We want to mark as Offensive the pages that are set up for the purposes of collecting pay-per-click revenue without providing much content of their own. You will see such cases most frequently in conjunction with “search results” feeds. Please read the whole section.

What you'll see on the result page: Usually, the page presents its own set of search results. Or, the page may look like the top-level page of a legitimate directory (tree structure) but clicking on a few selections reveals ads disguised as results. Or, you see copied content from a legitimate, credible resource, without value added by the copying site, plus a PPC program in place.

What's probably going on: The owner of the site gets paid whenever users click on these secondary results. You may be able to reveal this pay-per-click scheme by pointing your cursor to secondary links without clicking on them. Observe the status bar and you may see that clicks go through espotting, overture, or another advertising company.

Let us take a look at an example:

<http://www.startcool.de/Dir/Medien/Fernsehen>

This site is simply a copy of the Open Directory Project (aka DMOZ), but has a PPC program on the right (Google AdSense); the presence of AdSense PPC on top of the ODP content makes this site (every page on it) Offensive. Think about what the incentives are for creating a copy of the Open Directory Project; ODP is a free resource that does not accept advertising. By copying the search feed of DMOZ, sites can get contextual advertising on a pay-per-click basis. Google does not encourage creation of duplicates, so we are asking you to mark such result Offensive. Of course, had the result been a page on the Open Directory itself, it would have to be rated on the merits to the query. As you see, pages with the same content may be assigned vastly different ratings based on the absence or presence of a ppc program.

Here is an example of a page with ‘search results’ (ads):

<http://www.toxiclemon.co.uk/s.php?av=custom&ver=27617&set=uk-only&qkw=lastminute&qcat=web>

Note that the links on the page go through go2net.com. Also note: some ‘search result’ pages disguise the nature of what they do more than others. On Toxic Lemon pages, a more experienced user realizes that the results are essentially ads (Overture, Espotting are known providers of contextual ads), but this does not salvage the rating for this page. You can safely label all pages from Toxic Lemon Offensive, even if they are in another language.

Standard directories, or sites with results links that neither go through affiliate PPC programs nor redirect you through one of those programs, are usually not Offensive. One example of a non-Offensive directory is a directory that is clearly built by the site itself, not copied (<http://www.joeant.com/DIR/info/get/5704/48827>); also, a directory that charges for membership, not for clicks, is not Offensive. Consider for instance a directory of realtors that accepts entries for a yearly fee.

Please note that when you hover the cursor over links on the page you are examining, you are not always seeing the “true” URL in the status bar below. This is because it is possible to fool users by rewriting the URL reported in the status bar using Javascript, so take some extra time to understand where the links on the page are taking you.⁴

³ ODP (DMOZ) results are not Erroneous.

⁴ If you use Mozilla, you may have access to extra tools for spam evaluation. Write to us for specific instructions, please.

Some common PPC and Search Engine feed domains:

searchfeed.com findwhat.com espotting.com overture.com go2net.com

More examples:

<http://www.toxiclemon.co.uk/t/fancy-dress-shops/angle-dress-fancy-little-shop.htm>

<http://www.widgets.ws/widgets/us+robotics+modems>

<http://www.skc-networks.com/search.php?keywords=1260%20free%20nokia%20ri>

<http://hockey-apparel.discgolfnet.com/tennessee-titans-super-bowl-screen-saver.html>

<http://www.investment-wonder.com/top-search/investment/Group-Investment-Susquehanna.htm>

<http://www.carzilla.us/cgi-bin/search/search.cgi?keywords=motorcycle+rally>

<http://www.paley.com/search/Washing%20Machines.html> Clicking on ‘results’ on this page takes the user through affiliate.espotting.com; scroll to the bottom of the page on

<http://www.espotting.com/affiliate/account/login.asp> and you will see that Espotting.com engages in exclusive pay-per-click partnerships with European sites. Another example:

<http://www.shopguide.co.uk/Washing-Machines/>

<http://www.milipics.com/>

http://www.tools-directory.us/dir/matsushita_compressor/index.shtml

Thin Affiliate Doorway Pages

We differentiate between affiliates that produce extra service, value, or content, and those that simply are duplicates of other sites, set up to boost traffic to other sites and earn a commission for it. The former ones are not Offensive and should be rated on the merits to the query. The latter ones are Offensive. Please read the whole section.

UPDATE Please read Appendix I at the end of the Guide. Appendix I applies the distinction between thin content and added value affiliates to the case of Hotel Booking Sites.

Thin affiliate doorways are sites that usher people to a number of Affiliate programs, earning a **commission** for doing so, while providing **little or no value-added content or service** to the user.⁵ A site certainly has the right to try to earn income; we're attempting to identify sites that do nothing but act as a commission-earning middleman.

Observe where the links on the site take you. If the links are overwhelmingly leading you to one affiliate program, this is a strong signal that the site is a Thin Affiliate. Likewise, if the pages on the site are homogenous, and the links go to one or more affiliate programs, this is also a strong candidate.

In assessing sites for a Thin Affiliate rating, it is urged to click around the site (preferably during a "Sanity Check" in another browser) to determine if the links are affiliate in nature (or Pay-Per-Click, in the section that follows).

Here is an example of a Thin Affiliate:

<http://diesel-shoes.01shoes.com/Diesel-Mens-Retro-Shoes.htm>

This page has a number of marketing snippets for individual shoes, and a "More Information" button. Clicking on More Information button launches a popup window that takes you first through qksrv.net (Commission Junction), then to zappos.com. Zappos is known to have an affiliate program.

Clicking around the various navigational links on 01shoes.com shows more of the same design: a picture, a marketing snippet, and the link to Zappos via the Commission Junction; so, the correct rating is "Thin Affiliate."

The qksrv.net redirect is important to note, because online merchants often use a third party

affiliate provider to take care of the link tracking and payment. Thus the presence of these domains in the links on a page, or in redirects, can strongly suggest a Thin Affiliate classification:

qksrv.net
bfast.com
myaffiliateprogram.com,
webmasterplan.de
zanox-affiliate.de

Here's another example, this one using bfast:

<http://www.internetshopping.ws/1358.htm> ,

⁵ Usually the commission is not paid unless the user ultimately makes a purchase; contrast this with the pay-per-click schemes, discussed above.

Point your cursor to the link that says **Click here to buy ...** and observe the status bar window on the bottom of your window: you will see "http://service.bfast.com/bfast/click"

<http://www.internetshopping.ws/1367.htm>

<http://www.internetshopping.ws/1362.htm>

The www.internetshopping.ws site has nothing but affiliate links: no content, no service to users.

The following is an example of a site that was built using the Amazon API.

<http://us.store-directory.org/dvd/movie/B00005JM5E.html>

Note that all of the exits on the site for buying the product lead to Amazon. All of the content on the product page, including reviews, pricing, release dates etc. are available as part of the feed. The site adds nothing to the content that can be found on Amazon; it has no content value, nor does it add any service value to the user. A Thin Affiliate.

Here is an example of a site that should **not** be labeled Thin Affiliate:

<http://www.bookfinder4u.com/detail/0767914104.html>

At first cut it may look like yet another thin affiliate doorway to Amazon or B&N, but bookfinder4u.com is providing a **value-added service** to visitors by offering a comparison of prices between different online merchants. Ultimately you will be taken to Ecampus.com, Half.com, Amazon or another affiliate online bookseller, but the fact that they have their own price comparison infrastructure is the differentiator. To appreciate the difference, ask yourself this question: would any user **want** to go to www.bookfinder4u.com rather than directly to Barnes & Noble? To <http://us.store-directory.org/dvd/movie/B00005JM5E.html> rather than to

Amazon? The answer to the former question is Yes, because at Barnes & Noble, the user would not be able to see any direct price comparison between the B&N's price and competitors' prices for any given item; the answer to the latter question is No or Indifferent between the two. Surely, most naïve users may not even be aware when they are redirected, thrown from one site to another, etc. But if they were advised of what is going on, would then make an informed choice to go to a totally thin, no-unique-content affiliate doorway?

Another example of a page that does not fit the criteria of affiliate spam:

<http://www.mothering.com/books/books.shtml#adoption> gives a list of links that all lead to Powells.com, an on-line bookseller site. Clearly the Mothering magazine earns something when the readers buy books from Powells; however, equally clearly, the page is not set up for the sole purpose of generating affiliate links: browse the site a bit and you will discover that it has rich contents. Do not call a page affiliate spam when an affiliation is only incidental to the message and purpose of a website. To determine whether participation in affiliate programs is central or incidental to the site's existence, ask yourself this question: **Would this site remain a coherent whole if the pages leading to the affiliate were taken away?**

Another Example: http://books.webwab.com/item_512913.htm (clicking around on that site, you'll realize that every page simply leads to overstock.com pages)

More Examples: <http://www.thenewwidgetsite.com/prod/Kitchen-Etc/3-M-Command-Adhesive-Designer-Small-Hookss{1}Pack-of-2.html> PPC and A Thin Affiliate; a spam page with evidence of *multiple* spamming techniques is not a rare exception. <http://www.computermonitoruk.co.uk/>

<http://www.malls.cheap-money.com/>

<http://www.mabuy.com/News--Politics-magazines/The-New-Yorker.asp> - a doorway to Amazon and to Ebay.

At times the result page does not fall under any of the above categories yet still strikes you as “fishy”. In those cases we invite you to run the query on Google setting your preferences to show the top 20 results. View the first result page and try to find the URL you are rating. (You won't always be able to, as the result sets may have changed). If it is not in the current top 20, please rate the questionable result on the utility scale and move on. If it is in the top 20, examine the result set observing, where available, the following features:

- Do most of the top results resemble each other, and the result you are rating, in the snippets, titles, and/or URL structure?
- Do the result pages, when you click on them, resemble the result page you are rating in content? Contact information? Nearly identical, templated design? Affiliation with the same commercial entity?
- What about the snippets for your URL? Do they contain dictionary-like lists of words? Repeated text?

If your answer to several of the questions above is Yes, please rate the suspicious result as Offensive. If suspicion seems unjustified – all checks come out negative – please do not give the

Offensive rating. Not sure? One attribute, for example repeated text in the snippet, may or may not be a spam signal. So, send a question!

6

<http://www.google.com/preferences?q=gf&hl=en&lr=&ie=UTF-8&oe=UTF-8> Go to Number of Results and set to Display 20 results per page.

UPDATE Appendix I.

Hotel booking sites: spam or not?

Rating hotel booking sites is not easy. The technical questions – is it a real agency or just an affiliate? – has to be balanced against the user value considerations. We will address this issue now by giving examples of what is and is not spam.

First off, be more stringent when hotel booking sites come up as a result to a location query than when they come up to a hotel query. In other words, if you are dealing with a borderline case, resolve your doubts in favor of the Offensive rating if the query is for a location. Why? It is especially undesirable to have hotel booking sites crop up to queries that might presuppose hotels in the location of search, but might also look for a million other aspects of the location, such as reviews, transportation, a municipal site, a good resource on local history and geography, and the like. In a borderline case to a clearly hotel query (examples of such a query: [holiday inn, Cortland], [crowne plaza northstar hotel minneapolis], [Boston Park Plaza]), you may be more lenient. This is because the user intent is more unequivocally to get information on the hotel of choice, or to get a list of hotels in a location of choice. It can be argued that an opportunity to get a good deal on booking, the opportunity that some of the sites offer, is enough to warrant a merit-based rating for a hotel site.

Further, since there are affiliates and affiliates, it is important to differentiate between those who provide value added and those that just copy content and features off a feed to gain affiliate revenue without investing in offering unique and helpful services for the users.

As a fine example of the former, value-added sites, consider

<http://www.europeforvisitors.com/europe/articles/amsterdam-3-star-hotels.htm>

This site has a wealth of original articles (just do a few quick clicks around). Granted, most of the links on the above URL go through venere.com to get booking revenue, but the site as a whole offers a lot more than just stock hotel descriptions and booking links. Also, the comments and the apparent hand-selection of links is a definite value added service by the webmaster.

[holiday inn, Cortland],

http://traveldeals.sidestep.com/Hotel_Deals/New_York/Cortland/All/Holiday_Inn_Cortland?tk=EIKTDHHPXXX0000002 This site offers the users a download of an application to compare prices side-by-side and search travel sites. Not all users will find the application trustworthy, or worthy the extra time in learning how to use it in general, but we clearly do have an added service here – it's not just the same content off a feed. Hence, rate on the merits (Relevant).

[Boston Park Plaza]

Vital: <http://www.bostonparkplaza.com/> or

<http://www.bostonparkplaza.com/default.asp?SID=home> (remember, duplicates get the same rating) Not all hotels have their own homepages; for those that do, be sure to identify the uniquely authoritative nature of those pages by giving them Vital or Useful (as the case may be) ratings. It is sometimes difficult to do the differentiation because you see the same images on the official site, on the site of true travel agents, and finally, on the multiple affiliate sites...

Let us now walk through a handful of results to this query and make a determination on spam versus relevance rating.

<http://www.reservation-services.com/bostonparkplaza.html>

What immediately strikes as unusual is the candor in the disclaimer: “The telephone number, fax number and email addresses on this site DO NOT connect to the hotel.” Many affiliate sites list contact information right under the name of a hotel, so that users may be under the impression that they can call the hotel direct. Also, this site has its own staff: http://www.reservation-services.com/about_us.html; the names of the management staff are provided. This is a piece of evidence in favor of merit-based rating: this is not just a site that is set up as a middleman between the customer and the true reservation site. Finally, prominently behind the logo you see the link to “Become An Affiliate” – follow the link and see the offer the site makes to hotels. Clearly the site acts as a travel agent between the customer and the hotel, not as an affiliate of another booking site. So we are almost ready to give a relevance rating... but wait, let us go back to <http://www.reservation-services.com/bostonparkplaza.html> and check for hidden text. Sure enough, a few hidden keywords just below the copyright statement. Offensive. Find a few other hotels on this site and check for hidden text – you will see the same keyword white-on-white under the copyright statement.

<http://boston.hotelguide.net/data/h100012.htm>

Initially seems a borderline case. You see ppc (AdSense) on the right frame. You also see links to other sites bundled together: MetroGuide, EventGuide, DiningGuide, etc. (left frame); clicking on the first three displays information specific to Boston, so availability of these sites can be considered a value added. Nice to have also: links to local restaurants and nightlife. Are they an affiliate though? Yes; try to book and you will land on

<https://www.180096hotel.com/cgi-bin/bookit?SID=HG8&Dest=BOS&LKF=HGD&LANG=en&PROD=HOTEL+&DispCurr=USD&ITRK=dbP&qKey=YO330518800604&HtlId=NC+PARKP&Smk=N&Screen=0>

www.180096hotel.com, travelnow.com, ian.com and hotels.com are all one group. So you see that hotelguide.com has NO booking capability on its own and is signed up as a travelnow affiliate. And yet it is not spam. Why? It offers a video, an unusual and valuable added service. It subscribes to a travel video library <http://www.travelago.com/> to get additional content and service. This is enough to salvage <http://boston.hotelguide.net/data/h100012.htm> from the Offensive classification. Please rate on the merits to the query, taking into account that a video might make this site more helpful than other similar ones.

To reiterate: the added value provided by, first and foremost, the video, and also

http://travel.yahoo.com/p-hotel-397998-the_boston_park_plaza_hotel-i This is to remind you that special service pages that Yahoo provides, such as Movies, Finance, Travel, and others, should always be rated based on the merits to the query and not as Erroneous (of course not as Offensive either). In your merit rating, consider how helpful independent reviews by others might be to those who plan their voyages:

“Old gross bathrooms, stained carpet, chipped paint on walls, moulding falling off of walls, radiator falling off of wall. Room was the size of a dorm room. The only good thing going for this place is the location. We paid about \$190.00 a night - NOT WORTH IT!!”

D) <http://www.boston.the-hotels.com/boston-park-plaza-and-towers.htm> You see lots of links to other hotels. These seem placed for search engine spiders, not human visitors. The goal of the site is to get all of the hotels indexed. Evidence of spam. Pictures are nice, but where do they come from? Check the properties of any image and you will see they come from travelnow (an example: http://images.travelnow.com/hotels/thumbs/NC_PARKP-rooms-1-thumb.jpg). Let us try checking rates and we get to travelnow right away (<http://www.travelnow.com/hotels/hotelinfo.jsp?cid=46844&ID=122147>); so this is an affiliate of travelnow that adds no value, presents the feed available by signing up as a travelnow affiliate with nothing else. Images come as part of the feed. Spam – Offensive.

E) These two are not spam:

<http://travel.ian.com/hotels/hotelinfo.jsp?cid=54608&hotelID=122147&city=Boston&stateProvince=MA&country=US> and

http://www.hotels.com/best_hotels/us/ma/boston/boston_park_plaza_and_towers.jsp Ian.com (and hotels.com with its Benny the Bellhop logo, and travelnow.com) are a group that does the reservations (see <https://www.travelnow.com/itinerary/reserve.jsp?cid=46844>). They spawn affiliates but are not affiliates themselves (a critical distinction). Whitelist them, please.

<https://www.travelnow.com/itinerary/reserve.jsp?cid=46844> Tripadvisor, as you know, is whitelisted for the added value it provides in the form of reviews and rate comparisons.

<http://boston.guide-to-hotels.com/boston-park-plaza-and-towers-hotel.html> Again see a link to popular hotels by cities: Las Vegas, New York, etc. Cannot be there for the user (you usually are intent on going to Boston when you search for [boston park plaza] and not anywhere else) so must be placed for the spider. Is this site getting all content from an affiliate feed? Let us try sending a piece of the snippet to Google: [“The Boston Park Plaza and Towers is a traditional, landmark hotel”]

Sure enough,

<http://travel.ian.com/hotels/hotelinfo.jsp?cid=54608&hotelID=122147&city=Boston&stateProvince=MA&country=US> displays the same snippet. This is where the content comes from (you can see in Google listing to the search many more affiliate pages with identical information. Let us try booking on <http://boston.guide-to-hotels.com/boston-park-plaza-and-towers-hotel.html> And we immediately land on www.180096hotel.com: http://www.180096hotel.com/cgi-bin/chkrates?SID=BIO&Dest=BOS&LKF=BIO&TRK=_B4_link&PROD=HOTEL&Month=05&Day=29&Year=04&Nights=02&Adults=02&Children=00&Beds=1&Smoking=&LANG= So the content is off a feed, the reservations are through travelnow, is there anything added? Rating information, may be? In fact, the feed gives the rating information for the affiliates themselves, as a confidence index of the hotel's promptness in remitting the affiliate fee to the affiliate sites; it has nothing to do whatsoever with the guest satisfaction level.

Finally, if the site offers others to become its affiliates, it cannot be an affiliate itself. For instance, on the now whitelisted site www.180096hotel.com, notice the link to “[Affiliate With Us](#)”:

<http://www.180096hotel.com/cgi-bin/hotelinfo?SID=R10&LKF=MT4&HotelId=NC+PARKP&HName=BOSTON+PARK+PLAZA+AND+TOWERS&Dest=BOS&displayAd=false>

One cannot both be an affiliate of others and offer affiliation opportunities. So the presence of the link to become an affiliate is your hint that the site has its own booking functionality and can complete transactions for its visitors.