BlogBooker

From Blog to Book.

DEJANSEO.COM.AU/BLOG
Contents

1 2009

1.1 April ................................................................. 7
  1.1.1 Viral Marketing (2009-04-02 10:05) - admin .................. 7
1.2 June ........................................................................ 10
  1.2.1 NoFollow and PageRank Sculpting (2009-06-04 10:38) - admin ........................................................................ 10
1.3 December ................................................................. 10
  1.3.1 Micro-niche Search Results (2009-12-28 11:16) - admin ........................................................................ 10

2 2010

2.1 June ................................................................. 13
  2.1.1 Changes With NOFOLLOW tag (2010-06-04 09:18) - admin 13
2.2 August ........................................................................ 14
  2.2.1 Dropping URL Parameters (2010-08-06 10:31) - admin ........................................................................ 14
  2.2.2 The Art of Link Building (2010-08-06 10:42) - admin ........................................................................ 15
  2.2.3 Under the Google Radar (2010-08-06 10:45) - admin ........................................................................ 18
  2.2.4 Real Time Search (2010-08-06 11:15) - admin ........................................................................ 19
  2.2.5 Google Result Streaming (2010-08-26 09:07) - admin ........................................................................ 21
2.3 September ................................................................. 22
  2.3.1 Advanced PageRank Analysis by Dan Petrovic (2010-09-05 01:59) - admin ........................................................................ 22
  2.3.2 Correlation Between Latent Dirichlet Allocation (LDA) and Google Rankings (2010-09-09 10:51) - admin ........................................................................ 22
2.4 November ................................................................. 25
  2.4.1 Aussie SEOs Love to File Spam Reports to Google (2010-11-04 22:21) - admin ........................................................................ 25
  2.4.2 Google Image Search Still Lagging Behind (2010-11-06 20:53) - admin ........................................................................ 25
  2.4.3 Downward Spiral of Questionable SEO Practices and Why White Hat SEO Struggles to Survive (2010-11-13 17:32) - admin ........................................................................ 27
2.5 December ................................................................. 29
  2.5.1 Competitive Analysis of Australian SEO Results (2010-12-11 19:31) - admin ........................................................................ 29
  2.5.2 Practical Guide to Viral Marketing (2010-12-20 23:27) - admin ........................................................................ 51
  2.5.3 The End of Google PageRank (2010-12-31 18:15) - admin ........................................................................ 54

©2013 dejanseo.com.au/blog
Chapter 1

2009

1.1 April

1.1.1 Viral Marketing (2009-04-02 10:05) - admin

What Our Viral Marketing Services Can Do For You

When it comes to marketing yourself on the web, SEO is a good starting point, but to give yourself a little something extra, you might consider going with viral marketing.

What Is Viral Marketing?

People throw this phrase around quite a bit these days, and it gets misused a lot. So, to clarify, let’s give it a simple definition: viral marketing is an attempt at creating an advertisement that will be passed through the internet via word of mouth.

Notice that this definition didn’t include words like “funny” or “video” or “youtube”. Viral marketing can take place in any format, in any medium, on any website. All the word “viral” really implies is that the content “catches” like a virus, being passed from one person to another throughout the web.

Read more about

Brand and Product Identity

There are some scenarios where viral marketing is much more effective. In general, even the best viral content can take some time to catch on, so it’s really more suited for long term campaigns. If you’re holding, say, a one day sale or something along those lines, viral content may be too slow catching on to really be worth the effort.

On the other hand, brand awareness is really a perfect fit for viral marketing. Product launches can also be handled very well by viral marketing campaigns so long as you intend to keep this product around for awhile with no major changes in the near future.

Link building campaigns through viral marketing can be very effective, as well, as you’re simply building a larger presence on the web. At the end of the day, it’s really all about making a bigger name for yourself on
the web. As such, viral marketing can be incredibly effective for anything where you need long term public awareness of your business, products or services.

Creating Viral Content

What all viral content has in common is a "hook" of sorts. This hook can be politically charged, controversial content, it can be sexualized, humorous or dark content, it can even be a cute animal falling asleep. In any event, it doesn’t just transmit information, it transmits information in a gripping way, a way that people respond to immediately.

To take some case examples of the kind of content that can be incredibly effective in a viral campaign, you need look no farther than "Giant Stars".

This was an experiment in deliberate viral marketing conducted after a lot of research and planning. The aim was to create a video with a great hook and one that fell into a category and style that was already immensely popular on the web, that being the "world’s biggest..." "world’s ugliest..." "world’s weirdest..." etcetera. "Blankest blanks", for lack of a better name for the category.

The hook was the notion that however large you may think this or that star or planet is, there’s one that’s bigger. The video shows that the sun is millions of times as large as the Earth, while there are stars billions of times as the sun, and other stars billions of times as large as those stars.

This taps directly into a very basic human concept of our place in the universe. The video gets you thinking, it has a real "wow" factor to it, and it drives a fascinating point home: not only are we pretty small in the grand scheme of things, the things we consider huge are pretty small, too.

After being distributed through a few social networking sites, the video continues to gather thousands upon thousands of views on a regular basis and brought a [1]wonderful flow of traffic to the site within a few days.

Viral Awareness

Viral marketing is chaotic and unpredictable in more than a few ways, but it is possible to at least aim a piece of marketing towards going viral. For many, attempts at viral marketing can be incredibly hit and miss. Building a piece around a central, core concept, a hook, is a good start, but one should also be aware of what really appeals to the kind of people who pass these things on.

People who are skilled at creating viral content are those who are willing to sift through chain emails, portal sites and terrible content for hours on end looking for the universal concepts and conceits that all hot viral content seems to have in common.

From there, it comes down to quality.

Check our viral tests beneath:
Quality Content

A big part of what makes a video or a website or a song or an image go viral is quality, production value. The Giant Stars video features some basic, but professional looking computer imagery to show the scale of one planet or star to the next.

Something you hear emphasized in talk on search engine optimization is quality content, good writing. The reason some articles may get a couple thousands of reads while others, with the exact same SEO rating, will get tens of thousands is quality.

It’s easy to forget, in marketing, that you’re marketing to people, you’re not just trying to collect views, you’re not just trying to improve rankings, you’re trying to talk to people, you’re trying to share a product or service with them, make them aware of something.

Content, in search engine optimization or in viral content, is all about quality. Visibility is all about quality. Without quality, you’re essentially left collecting every view one by one, posting the video or article or song in one place at a time. Quality is what keeps people trading something back and forth, sharing it with one another on message boards and in chat rooms.

Creating quality content comes down to passion, in a sense. It’s said of screenwriting that if nobody laughs, cries or feels touched while writing a movie’s script, then nobody will laugh, cry or feel touched while watching the finished film. The same goes for passion in creating viral content.

In creating something that’s supposed to be funny, you need only put it to your own standards of humor: did you laugh at this video or article or song? In creating something like Giant Stars, the first person that has to say “Wow” is the person who creates the video.

What This All Means...

All we’ve really been discussing in the above text is how viral content works. When it comes down to actually creating a viral marketing campaign, a lot has to be determined between the content creators and client as the process goes.

Viral content has to be geared specifically towards a certain demographic, it has to be built around specific products or services that are being made available. Discussing your project with a professional who has
created and managed successful viral marketing campaigns in the past can, in turn, inspire several new ideas and can show you how closely tied together marketing and business really are.

Some go so far as to argue that the very best products are those that are conceived hand in hand with their marketing campaigns. Something like Pokemon, for instance. The top selling video game series of all time. The game itself requires children to play with one another to collect all of the in-game Pokemon. This is a real-life example of viral awareness: children are seen playing together at recess, and other kids, curious as to what’s going on, observe and want the game for themselves.

Strong viral marketing is all about creating that kind of perpetual awareness in the product or service. An awareness that spreads, like a virus, from one person to another and becomes impossible to ignore. Whether this is something inherent to the product itself or a certain way of seeing the product that the viral marketing suggests, the ultimate aim of a viral campaign is to, essentially, let the market do the advertising for you, and a well designed approach crafted by experienced viral marketers can accomplish just that.

1. http://www.giantstars.de/gs-stats.gif

1.2 June

1.2.1 NoFollow and PageRank Sculpting (2009-06-04 10:38) - admin

"...Matt Cutts dropped a bomb shell that it no longer works to help flow more PageRank to the unblocked pages." via SELands/ SEOMoz

If your SEO placed a nofollow tag on all your low relevance pages simply to gain better link juice in order to rank higher - this won’t work anymore according to the statement Matt Cutts made at SMX yesterday. Matt is the official Google spokesman for SEO community. An example would be to block privacy policy, login page, terms and conditions, contact us and similar pages in hope to preserve link juice for the product and service pages we wish to see in the result pages.

I my opinion Google did this due to the fact that too many valuable pages started to hide all this valuable information which was no longer in Google’s index, and that is of course a big concern for the leading search company. Another factor is that, nofollow tag should be used to prevent associate ion with pages you cannot trust which in many cases rendered someone’s privacy policy a statement the site itself does not trust which is a bit of a problem.

If you need help understanding the latest changes [1]contact us today.


1.3 December

1.3.1 Micro-niche Search Results (2009-12-28 11:16) - admin

[1]SEO News:
Google and Micro-niche Search Results

Due to some recent changes to Google’s search engine, many internet users may have noticed some unusual search results in recent days. This is due to a major change that Google has implemented in their search engine. While Google used to only take a person’s web history into account if they were logged into their account, Google now customizes all search results based on anonymous search history, as well as the location of their IP address.

This means that every single internet user will be receiving different search results, which begs the question of whether some of the current web marketing techniques will be rendered obsolete. Instead, it turns out that these changes have simply magnified the impact of any traditional marketing campaign designed to attract specific customers. The key point to understand here is the difference between web history and search history. When a user is logged into Google, his search engine results are based on many aspects of his behavior on the internet, ranging from sites that he has visited to developing trends in his recent web activities.

A person’s search history, on the other hand, is a record of the web searches that a person has entered into a search engine in the recent past. Google also takes the location of an user’s IP address for certain queries such as local restaurants and movie theaters. Regardless of whether or not an individual is logged into Google, every person will now be receiving different search results based on how they use the web. While these recent developments will have a big impact on internet marketing, it is important to understand that this does not change the game of search engine optimization in the slightest. Instead, customized search results simply maximize the impact of every SEO campaign that is designed to reach your core customers. In fact, these recent developments are great news for any business owner or web developer who is concerned with attracting and retaining a very specific set of clientele.

Because an individual’s web history and search patterns affect his search results both so heavily now, it is now more important than ever to make the prime goal of your SEO strategy to match the interests of your core demographics. Because Google bases their search results in part on an individuals web search history, the importance of implementing an intelligent SEO campaign is more important now than ever. So long as your sites already have a respectable ranking in the key search engine results for your demographic, these new changes do not require most web developers or business owners to make major changes to their present SEO strategies. Instead, Google’s new approach to customizing their search engine results to an user’s web history will simply compound the impact of a site that has implemented the correct SEO tactics. In addition to an individual’s search history, the new changes to Google’s approach to customized search engine results are affect by an individual’s behavior on the internet.
For instance, if an individual reaches your website through clicking on a banner ad or a list of links located on an associate’s website, Google will take this information into account the next time that the individual makes a search query on a topic that is related to your site. Even if this particular user had never reached your website through Google before, your site will rank near the top of the search engine rankings based on the presence of relevant cookies on the individual’s web browser. Again, the influence of link building is simply magnified by the influence of one’s web history on future search engine results.

Update:
When possible, Google will customize your search results based on location and/or recent search activity. Additionally, when you’re signed in to your Google Account, you may see even more relevant, useful results based on your [2] web history. The following information was used to improve your search results for seo:

Search History Your search results have been customized using search activity from this browser. [3] Learn more [4] Disable customizations based on search activity If you’re curious, you can see what a search for seo looks like [5] without these improvements.

(Comment by Dejan SEO: Variable added to URL is “&pws=0”) The ‘More details’ link on your search results page can be used to display this page for approximately 30 minutes, after which it will no longer show this page.

Chapter 2

2010

2.1 June

2.1.1 Changes With NOFOLLOW tag (2010-06-04 09:18) - admin

Ambiguous Role of the NoFollow Tag

5 December 2011 Update: In a recent Live Q &A session Matt Cutts states that they generally don’t follow the nofollow links. This is confusing as we all know that they do, at least for document discovery if not for passing PageRank. The question is did Matt Cutts mean the latter, or has something changed at Google and nofollow links are truly not followed by bots anymore?


NoFollow tag and PageRank Sculpting

4 June 2009 "...Matt Cutts dropped a bomb shell that it no longer works to help flow more PageRank to the unblocked pages.” via SELands/ SEOMoz If your SEO placed a nofollow tag on all your low relevance pages simply to gain better link juice in order to rank higher - this won’t work anymore according to the statement Matt Cutts made at SMX yesterday.

Matt is the official Google spokesman for SEO community. An example would be to block privacy policy, login page, terms and conditions, contact us and similar pages in hope to preserve link juice for the product and service pages we wish to see in the result pages. I my opinion Google did this due to the fact that too many valuable pages started to hide all this valuable information which was no longer in Google’s index, and that is of course a big concern for the leading search company.
Another factor is that, nofollow tag should be used to prevent associate ion with pages you cannot trust which in many cases rendered someone’s privacy policy a statement the site itself does not trust which is a bit of a problem. If you need help understanding the latest changes [2]contact us today.

1. http://www.youtube.com/embed/R7Yv6DzHBvE?rel=0

2.2 August

2.2.1 Dropping URL Parameters (2010-08-06 10:31) - admin

Google Tests Canonical Page Versions by Dropping URL Parameters

SEO firms and webmasters are sure to be interested by the news revealed by Google’s Matt Cutts that was recently posted online. Responding to a question regarding whether or not the Googlebot uses inference when spidering the Internet, Cutts agreed that it definitely does - to some extent - and that it looks at redundant parameters in order to do so. What does this mean for SEO, webmasters and other interested people? It might hold a great deal of significance when it comes to using the rel=canonical tag that started coming into vogue in early 2009.

How Does Googlebot Handle Unnecessary Parameters?

Cutts explains that the Googlebot uses inference while spidering by looking for duplicate parameters, or for parameters that appear again and again; in other words, it looks for overly ubiquitous parameters. When it recognizes one as such, it attempts to drop it to see if it still achieves the same results. For example: www.example.com/index.php?page=pagename &parameter1=green &parameter2=widget might be tried out using the URL www.example.com/index.php?page=pagename &parameter1=green in order to see if the content is affected significantly. If it isn’t, Googlebot will drop the unnecessary parameter - in this case, ”parameter2=widget” - to create cleaner indexed URLs and to reduce the amount of redundant search results in the Google rankings.

How Does Googlebot’s Handling Of Parameters Affect The rel=canonical Tag?

SEO professionals and other web developers were understandably excited by the news about the rel=canonical tag, which allows users to specify which version of a page should be prioritized by Google. In the past, the process was determined through complex Google algorithms and was decidedly imperfect. Naturally, those algorithms are still at play, but developers can take more control by using the rel=canonical tag to let Googlebot know which version of a page should hold precedence. However, the news that Googlebot uses inference when spidering, and that it drops unnecessary parameters when considering which content is relevant means that the rel=canonical tag might be largely worthless. After all, if Google weeds out a parameter that would drop a huge number of pages from consideration - but your rel=canonical tag points to one of them - it’s not going to do you any good in the long run. This news
2.2. August

BlogBook

highlights the fact that SEO is, indeed, a true science. When crafting an SEO-friendly site, experts will have to take Googlebot’s inference practices and the rel=canonical tag’s relation to it into consideration. Whether or not the rel=canonical tag will ultimately be worth it remains to be seen, and will have to be determined on a case by case basis.

2.2.2 The Art of Link Building (2010-08-06 10:42) - admin

Link Building Services | Link Building Company

Our Link Building Strategies

Following are the link building strategies we employ in order to grow and maintain powerful and diverse backlink portfolios for our clients:

Standard Level Link Building

- Advanced Link Target Research, Discovery & Negotiation
- Competitor Backlink Acquisition
- Establishing Passive Link Building Tactics
- Link Consolidation and Non Anchor References
- Quality Directories
- Relevant Memberships
- High Value Donations
- News Publishing Sites
- Exploring Non-Competitive Industries
- Brand Leverage Tactics

Advanced Methods

- Link Bait
- Citing and References
- Viral Marketing
- Video Distribution
- Widgets
- Competitions
High Quality Link Acquisition & Consolidation

Dejan SEO link building team employs a set of unique link acquisition tactics that secure high quality links in addition to mass link generation typically encountered with traditional link building companies. Our strategy has been tested and applied to numerous campaigns in the past with great success. Our tests indicate the need for the right balance of high quality links within the mix of regular link building. Low level backlinks will not perform alone within a desired period of time. Summary: Quality links play essential role in getting quick results in search engines.

Quality Backlinks

Quality backlinks are defined by the following criteria: Google PageRank 4 or higher, being in DMOZ or Yahoo! Directory, part of an .edu, org.au or asn.au domain or having an authoritative status through high relevance, traffic or backlink portfolio. Our link building team is highly efficient at securing links which means that you will generally receive a link for every $100 or $200 of the campaign budget.

For example: Campaign of $8000 for example can secure 40 to 80 high quality permanent links. The success rate increases each month with higher familiarity with products, services and linking opportunities. Our link building team learns and self-improves focusing on most efficient ways of securing quality links.
2.2. August

Long Tail

The increase in rankings will be steady and gradual and designed with long-term effects in mind. We expect a significant portion of traffic coming through a long tail phrase range (niche terms) prior to flow of traffic from some of the more competitive target search terms.

Maintenance

Dejan SEO will ensure that a constant flow of fresh backlinks is continuously added to the site in order to maintain high rankings. Any phrase refinement, campaign adjustments and ongoing on-site recommendations are covered and handled by our consultants. We feature a high level of transparency and provide 24/7 access to rankings and link building activities.

Dejan @ SMX Sydney: Link Building Strategies

[EMBED]

Download: [1]Dejan Petrovic @ SMX, Sydney 2010 Powerpoint Presentation Slides

More About Link Building

Google and other search engines use numerous signals to determine website’s reputation and authority, however, the strongest factor out of all is the quality and diversity of incoming links from other websites. Essentially the whole logic is based on the voting principle where each link from a website other than your own is a vote. This might lead to a quick conclusion that more links you get the better, which can be a deceiving thought and later in this article we will find out why. Let’s say you’re looking for a good lawyer.

You are flipping the local business directory, seeing ads and hearing about different names and brands. Which one is better you ask yourself? This is where you ask your friends, colleagues and relatives. Some of people you know will be more trustworthy and some will be better acquainted with the law business whereas others advice will not weight as much due to their lack of expertise, experience or they are simply someone who you would not trust with advice. Same rules count for links pointing to your website. Search engines will look at your content first and try to understand what your website is about.

After that they will try to see if they should trust your content by seeing who links to you. But this is not where it ends. Similarly to the way people determine who to trust in the real world search engines will treat links from different website in different ways. For example, these are pretty important factors:

- Is the website that links to you related to your website’s content?
- Are others linking to the website that links to you?
- What is the quality of the links that are pointing to the website that links to you?
- Where are you linked from on their page?

Getting one link from a trusted source is better than a thousand links from spammy websites. A site owner can get pretty good links and rankings through simply generating outstanding content or providing excellent service. This alone, however, is not always enough as there will be always those who do that – and something
extra to achieve better rankings.

It’s this ‘extra’ that gives you the competitive advantage and this is the main benefit you will receive from your SEO company or a link building specialist. Before embracing a link building campaign we always ensure that we’re not basing our efforts on pure guesswork. Phrase research is of absolute importance for any link building or SEO campaign. Targeting wrong phrases could be a tremendous waste of time, money and resources. Another thing to watch out for is to make sure that your website doesn’t get into a ‘bad neighborhood’ so make sure you don’t participate in dodgy linking scheme and backhat SEO practices that are in conflict with search engine guidelines.

Our link building specialist team features years of experience in securing quality links that work well in line with your on-site optimization and content. If you would like to find out how your business can benefit from our advanced link building strategies please contact us or call [2]1300 123 736


2.2.3 Under the Google Radar (2010-08-06 10:45) - admin

Will Small Traffic Black Hat SEO Sites Stay Under The Radar?

Do you think you can beat the system by hiring a black hat SEO company? Do you think using sleight of hand will get you through the roadblocks set up by Google? If you do it might be time to think again. Black hat SEO companies will tell you they have a way to shimmy under Google’s radar. On the surface they aren’t lying. Bold faced lies, in fact, are not the way black-hatters operate. They prefer to evade the truth rather than stomp on it. It is what they are not telling you that will get you into trouble in the long run.

Short term black hat tactics can, indeed, work to get you past Google’s imposing algorithms. Black hat SEO companies prefer to set up multiple small websites that [1]fly just under Google’s radar because they are not considered an imposing threat. Then they use those sites to spam messages far and wide and believe they won’t get caught. Eventually the axe will fall.

One of the problems of black hat tactics is that in order to maintain a low profile the user must keep their site small and unvisited. Exactly what are those businesses paying the SEO company for then, if not to drive traffic to them? The entire idea behind SEO is to improve rankings, and get people (potential customers) to a website. By its very nature this scattered buckshot SEO tactic defeats itself. Even if two hundred individual sites are set up and used to spam for low-level traffic any good sales person will tell you it takes a much larger number of visitors to accumulate the needed sales for true success unless your product is an extremely high-end and expensive item.

Another critical error you make by using an SEO company that employs the buckshot SEO tactic of multiple small sites is that eventually your number will come up. Google does use human browsing and intervention along with their automated bots and spyders. Where the automated system may overlook your spam shots because of the relatively small size of your website, the individual will not.
Companies that use these tactics are hedging their bets that Google’s resources are not large enough to permit them to find every problem site. It will take time for Google to find them in most cases. They plan to move on as soon as they are discovered and open a different set of sites. This is very labor intensive if you think about it. The time you spend building a site and attracting legitimate consumers to your online shop is wasted once Google discovers it and you have to move on and start over.

1. [http://www.youtube.com/watch?v=fvlS12bJnF0](http://www.youtube.com/watch?v=fvlS12bJnF0)

2.2.4 Real Time Search (2010-08-06 11:15) - admin

Zero to First Page in 10 Seconds

Dejan SEO, 29 December 2009 It was many years ago that one had to "submit to search engines" and today it is enough to just get a single link or link out and your existence is known to most search engines. Pages have been entering Google index even without any active announcements via certain levels of inference and other signals such as syndication, sitemaps, references and various other tools. One thing still troubles SEO community - speed of indexing.

It is a well-known fact that high-traffic / high-authority websites get spidered all the time and their pages enter Google index within hours and even minutes. The fact of life is that we cannot all be treated as authority and a vast majority of websites remains in the slow lane waiting for page indexation for weeks and sometimes even months. Google didn’t seem to be in a rush to provide faster and more current results until the Twitter revolution. Twitter prompted Google’s latest algorithm update (Google Caffeine) and the introduction of a rather patched up version of real time search.

How to get into Google results in less than 10 seconds?

It’s simple. Just send a Twitter message by your phone, app or browser. If your message contains the target keyword, the odds are it will show up in Google "real time" search results as part of Google’s universal search result mix.

The process is illustrated here:

1) Dejan posts a message about the latest SEO package. See image below:
2) Five seconds later Google inserts our Twitter post into their "real time" search results:

How useful is this?

Well, for an active Twitter user this could bring a moderate exposure and branding, and even some traffic. For the rest it doesn’t bring much value. However we should all ask ourselves, can we allow to get cushy and comfortable with our search engine results? Consider the effect of the following on your visibility:
2.2. August

- Real-Time Search
- Increasing Speed of Indexing
- Personalisation
- Universal Search Results
- Semantic Web
- Search Suggestions

Search engine optimisation is starting to blend more with PR, marketing and advertising and 2010 will reinforce this even more with the introduction of new variables in the world of search engines.

2.2.5 Google Result Streaming (2010-08-26 09:07) - admin

Today an interesting new feature of web search giant Google has emerged - live results updating. Rob Ousby, a SEO consultant, broke the story when he noticed the phenomenon and posted it to his blog.

Since then, a number of other news organizations have picked up the story. Most notably, TechCrunch not only ran the story but was also able to confirm with Google that the video is real. The new feature, to avoid confusion, is that Google is actually updating the results of a page, not just search suggestions. That means that as a user types in their query at the top of the results page, the page below them is continually changing and updating. This could potentially save users lots of time typing and finding results, as they may not even have to finish a search query before finding what they are looking for.

Interestingly, though, is that while this seems like a great new feature, the user response doesn’t seem to be great. Scanning the comments on Ousby’s blog reveals a myriad of complaints from users who feel that the new feature is too flashy and will be annoying. Several commenters suggest that it will simply be too distracting and hinder them form typing in their complete query. Thankfully for those users, Google hasn’t actually launched the new feature yet. It is currently just a test, which is being rolled out to a small portion of users. Google has said that at any one time, they are running anywhere from 50 to 200 tests, which are only seen by a very small percent of their users. Google is known for making tiny tweaks to their pages, which can make huge performance differences to users.
Small changes for instance, in how ads are displayed can make huge differences in Google’s revenue, and so they are constantly trying to improve their web search service. Chances are that you won’t see this, and if you do, then you are probably a rare part of one of Google’s current tests, and this feature may not stick. This could be seen launched to all users at some time in the future though.

1. [http://www.ousbey.com/blog/live-updating-google-search-results](http://www.ousbey.com/blog/live-updating-google-search-results)

2.3 September

2.3.1 Advanced PageRank Analysis by Dan Petrovic (2010-09-05 01:59) - admin

Recently, Google’s Matt Cutts talked about PageRank sculpting and mentioned something which shattered many people’s PageRank sculpting theories and practices. He explained that using the HTML nofollow attribute to prevent search engine web crawlers from going to unimportant parts of a website, like a privacy policy, does not give you a higher ranking on the other pages on your site.

Private testing of this seemed to confirm his claim. The addition of a nofollow tag to some links did not improve the PageRank of other pages linked to on a site, however the PageRank could have changed, just not by a whole number, so that the change wasn’t visible (such as going from 4.5 to 4.75).

Though that did seem to confirm Matt’s claim, another observation that was made didn’t. If you have 10 links on the home page of a site, they usually go down by one PageRank value (example: 5 to 4). Having 100 links on the home page meant that the pages they pointed to couldn’t get higher than a PageRank of 2 or 3. This contradicts Matt’s claim that having less links on a page doesn’t make the PageRank of those other pages go up, and that the number of links doesn’t make a difference at all. This makes no sense, and it would be great to get clarification from Matt and to hear his view on this issue.

The conclusion from all of this is that Google does, in fact, follow nofollow links, but they just don’t take them into consideration in determining a page’s PageRank. "No follow," therefore, is more like "no vote." It is worth bearing in mind, however, that Google uses many other factors to determine a PageRank, like on-site positioning, navigation depth, and off-site links and contextual references. This means that it is much harder to tell what factors are affecting a PageRank and that our job is that much harder.

The conclusion from all of this is that it is not worth all of the time to try to place internal nofollow tags, but rather to spend your time creating more content and generating more links to your site.

2.3.2 Correlation Between Latent Dirichlet Allocation (LDA) and Google Rankings (2010-09-09 10:51) - admin

A recently published article on SEOmoz’s Daily SEO Blog discusses an exciting new find in the field of [1]SEO - the discovery of high correlation between Latent Dirichlet Allocation (LDA) and Google’s search results. Basically, SEO researchers have found one method that Google uses to rank search results which we
previously didn’t know about, and we may be able to use this to optimize pages better. The article, written by Randfish details a talk given by Ben Hendrickson at SEOmoz’s recent annual "Mozinar".

LDA is a complex topic which we won’t go into fully. What you need to know to understand this post is that LDA is a type of “Topic Modelling” – a method of relating words to each other. The blog gives a great example of this in a diagram using the words "cat" and "dog" as an example.

![Simplistic Term Vector Model by SEO Moz](image)

Here, all words are related to either cat or dog. A neutral word, like "bigfoot," would be 45° away from both words. Words like "feline," or "canine" would be much closer to cat and dog, respectively, because of their increased association with each of those. This model is drastically oversimplified - in actual LDA, there would be millions of different words and phrases which all exist in separate dimensions.

However, for the sake of this post, understing this simple example is enough. Now, how does this apply to SEO? Researchers at SEOmoz investigated using an LDA modelling to replicate Google search results. We know that Google uses over 200 different factors to determine search results, and even though we know what some of these factors are, we do not know how they are all individually weighted. In the past, SEO
Researchers have used several different models to rank search results, and compared them to Google’s results. Using simple models like keyword frequency, or TF*IDF (term frequency x inverse document frequency), they have found some correlation with Google’s search rankings.

Correlation does not, of course, imply causation, but it does give strong suggestions about how much weight Google places on various factors. The big break through with this LDA model was its high correlation - a little over .33.

This may sound underwhelming to the average person, but this is significantly above the correlation shown by any of the other models. Even though this isn’t conclusive, it suggests that term modelling and LDA may be a major factoring used by Google to determine search relevance. One other pieces of evidence helps to back this up. Every two years, SEOmoz conducts an opinion survey of SEO professionals to try to determine the weight that Google places on various factors.

The most recent survey showed a major increase in on-page keyword usage. This also suggests that Google is using some kind of term modelling or LDA system to determine relevance. All this is great, but if you are an SEO professional, then you are probably still wondering how it impacts you or how you can use it to your advantage. To understand this, let’s go back to what term modelling is. Term modelling is based on figuring out how similar some words are to others.

Instead of thinking of this as keywords, think of it as keyword synonyms, or keyword keywords. In addition to using keywords on your page, you may also want to use words related to that word. For instance, a page which is being optimized for the keyword ”computer” should also contain words like monitor, CPU, and hard drive, but using the word CD might make the search engine think the page is about music. This new research suggests that using these additional keywords will help improve a page’s ranking for the original word.

Because Google uses a very complex search algorithm, and all of the 200 different factors that it uses are weighted differently, we will never know exactly what its search algorithm is. This research does not suggest that LDA is more than a small part of the algorithm, but it does show that it is significantly more important than was thought before. For more on some of the terms used in this article, check out these links:


Related: [6]International SEO Research (French)
2.4.1 Aussie SEOs Love to File Spam Reports to Google (2010-11-04 22:21) - admin

By analysing the raw SEO survey data supplied by [1]SEOMoz we found that 46 Australian SEOs submitted a spam report against another site. Funny thing is that the stats also show that those who report others are more likely to have also done something naughty. Why is this so amusing? Because these individuals are harming themselves in the process. How? Well, Google is a search engine that likes to do things algorithmically and with minimal human intervention (unlike some other search engines which have more of a human touch to it, for instance Yahoo and their directory).

After reviewing the feedback from these reports [...] it may take a bit of time before we update our algorithms and a change is visible in the search results.

Source: [2]Google

When you report a website to Google you help feed more data into their algorithm and aid them in future detection of spam and those who are trying to cheat the system. If you are also cheating the system you are in the same boat as the one you just reported. Human intervention is rare and when website gets penalised its usually due to an algorithmic trigger.

Like with everything else, Google tries to be smart about their resources and take action against sites whose actions affect most users and going down from there. We wrote something about how some small [3]websites stay under the radar for longer due to this effect.

Our advice to webmasters? Be smart about your SEO and earn search engine love through excellence in content and service, but if you are taking shortcuts to your success in rankings be prepared to do some explanation to Google when the penalty arrives.


2.4.2 Google Image Search Still Lagging Behind (2010-11-06 20:53) - admin

Ever since Bing and Twitter nudged the web towards real time search, Google has been trying to get it right, but it all seems a bit too rushed and unfinished. An example of that is evident in the universal search
segment. For example if we search for "[1]prime minister australia" in Google Image Search we get numerous photos of Kevin Rudd. This tells us that image algorithm differs from the mainstream search and it would be the right time for Google departments to consolidate and update across different platforms. By modifying the search to "[2]current australian prime minister" we get a slightly improved result, but still not good enough. After modifying the search term unsuccessfully numerous times we decided to type in a very direct search: "[3]australian prime minister gillard" and what do we see on the first result? Yup. Kevin Rudd again.

Google Image Search boasts numerous features such as size, type and colour filter which is great, but there is a need for time-based filter as well, similar to classic search:

- Any time
- Latest
- Past 24 hours
- Past week
- Past month
- Past year
- Custom range...

It might be worth mentioning that there is a connection between custom date search and image results ("[4]Updates with images") which seems like an excellent compromise, until you click on it and receive no results. Good news is, however that you can always expand the date range to say, [5]24 hours to harvest more results.

Great effort Google, but plenty more to do!


2.4. Downward Spiral of Questionable SEO Practices and Why White Hat SEO Struggles to Survive (2010-11-13 17:32) - admin

Online industry can be difficult for even the toughest guys on the block. One moment you think everything is going fine for you, you are doing your own thing, SEO your website by the book, starting to see progress, expecting that first page and eventually top ranks in the next few months. So what happens, while your website is entering its 3rd year here comes a 4 months old website sitting at the top of Google, wait, what?

Yeah, I know, it is mighty frustrating and it surely doesn’t help you see your goal clearly. I mean here you have a guy that clearly paid for links, used some Black Hat SEO techniques and what does he get; a penalty? Nope, he gets to open up a bigger bank account and light a cigar. Not only is a website not worthy of being number one at the top, but you got yourself another competitor and lost a rank, which means less traffic and fewer customers, so what do we do in that situation? We go and smash our piggy bank and go buy some links Dorothy, that’s what!

Yeah, we find ourselves a nice link dealer and get ourselves, 100, no, 200 high PR backlinks and what do you know, next week we are the prom queen of the Google homecoming ball. Ah, now it’s ok, we can breathe without worrying about our rankings, that guy actually did us a favor, if he didn’t buy links we would have never gone that way and look at us now, top of the world. Until the next morning when you wake up to see that you actually don’t see, sandboxed baby, Dorothy is gone and your website moved from OZ.

This actually happens all the time. We are bound and forced to respect Google’s rules as they are the largest search engine, the big boy on the hood making the rules of the playground. But what happens when Google ignores a website that used Black Hat SEO techniques and got to the number one place using paid links? Google is making it worse for all of us. I mean looking at that website I might be tempted to do the same, if he can get away with it so can I, right? Well Google is selective and although 9 of 10 website will get penalized for such behavior there will be that 1 website that got away, that beautiful homecoming prom queen everyone talks about.

It sure is a bait for all of us to try that road, but it probably won’t work, we are not that lucky. But I don’t blame people for trying, I mean with all the websites that got away Google is actually asking us to try, they complain about the amount of SPAM they receive and Matt Cutts says that it’s hard to deal with all the
SPAM on the net, well what can I say, I’m not the one who invited SPAM into my own playground, Google did.

Google is getting the amount of SPAM equal to the number of invitations they sent, so it’s their problem, not ours. But even though it is their problem we are the ones suffering. Going white hat can be a pain in the neck, it can take a long time to achieve your goals, but although it seems too hard at certain moments it is actually the only way to go, and if you know what you are doing or your SEO company, it can be fairly easy.

Content is king, you heard that plenty of times and you will hear it more and more as it is true. And since content is king everyone loves being seen with the king. It’s the rule of celebrities; one picture with a celebrity may make you famous. So if you have the king on your side people will link to you, saying hey, I know the king he is a good guy, go check this out and they point a link to the great content you created on your website.

The bottom line is that online industry is not all black and white; there is plenty of gray as well. It’s not fair that Google will let one website profit from paid links and punish the other, but this is a fact that we have to deal with. The only choice we have as webmasters is to play the White Hat SEO game and give our best to create valuable and quality content that people will want to link to. This is the only method that will get you links and links for years to come, paid links are there as long as you pay for them, but free links that no one asked for can stay in their place for a very long time.
2.5 December

2.5.1 Competitive Analysis of Australian SEO Results (2010-12-11 19:31) - admin

DejanSEO team has performed a competitive analysis across top 100 Google results for "SEO" and "seo company". The dataset included 689 words occurring in 100 different page titles with 160 unique words. Complexity factor (Lexical Density) was found to be 23.2 % and the Gunning-Fog Readability Index at 10.6 which puts it on the "easy" side of the readability scale. Total number of characters analysed was 5914 with and 4739 without spaces.

Top phrases used in titles are:

seo, search, engine, optimisation, company, services, marketing, melbourne, website, sydney, brisbane, australia, optimization, web, internet, design, experts, optimisation, australia

Full List of Phrases:

Table below shows all of the 160 words used and sorted by use frequency:

<table>
<thead>
<tr>
<th>Word</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>seo</td>
<td>29</td>
</tr>
<tr>
<td>search</td>
<td>41</td>
</tr>
<tr>
<td>engine</td>
<td>42</td>
</tr>
<tr>
<td>optimisation</td>
<td>38</td>
</tr>
<tr>
<td>company</td>
<td>32</td>
</tr>
<tr>
<td>services</td>
<td>24</td>
</tr>
<tr>
<td>marketing</td>
<td>21</td>
</tr>
<tr>
<td>melbourne</td>
<td>18</td>
</tr>
<tr>
<td>website</td>
<td>16</td>
</tr>
<tr>
<td>sydney</td>
<td>14</td>
</tr>
<tr>
<td>brisbane</td>
<td>12</td>
</tr>
<tr>
<td>australia</td>
<td>11</td>
</tr>
<tr>
<td>optimization</td>
<td>10</td>
</tr>
<tr>
<td>web</td>
<td>9</td>
</tr>
<tr>
<td>internet</td>
<td>8</td>
</tr>
<tr>
<td>design</td>
<td>7</td>
</tr>
<tr>
<td>experts</td>
<td>6</td>
</tr>
<tr>
<td>optimisation</td>
<td>5</td>
</tr>
<tr>
<td>australia</td>
<td>2</td>
</tr>
<tr>
<td>Keyword</td>
<td>Count</td>
</tr>
<tr>
<td>--------------</td>
<td>-------</td>
</tr>
<tr>
<td>seo</td>
<td>138</td>
</tr>
<tr>
<td>search</td>
<td>73</td>
</tr>
<tr>
<td>engine</td>
<td>73</td>
</tr>
<tr>
<td>optimisation</td>
<td>46</td>
</tr>
<tr>
<td>company</td>
<td>21</td>
</tr>
<tr>
<td>services</td>
<td>19</td>
</tr>
<tr>
<td>marketing</td>
<td>19</td>
</tr>
<tr>
<td>melbourne</td>
<td>17</td>
</tr>
<tr>
<td>website</td>
<td>12</td>
</tr>
</tbody>
</table>
2.5. December

BlogBook

1.7 %
7

sydney
12
1.7 %
7

brisbane
11
1.6 %
8

australia
11
1.6 %
8

optimization
10
1.5 %
9

web
9
1.3 %
10

internet
8
1.2 %
11

design
7
1 %
12

experts
6
0.9 %
13
<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>Percentage</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>optimisation</td>
<td>5</td>
<td>0.7%</td>
<td>14</td>
</tr>
<tr>
<td>australia</td>
<td>5</td>
<td>0.7%</td>
<td>14</td>
</tr>
<tr>
<td>seo</td>
<td>4</td>
<td>0.6%</td>
<td>15</td>
</tr>
<tr>
<td>professional</td>
<td>4</td>
<td>0.6%</td>
<td>15</td>
</tr>
<tr>
<td>agency</td>
<td>3</td>
<td>0.4%</td>
<td>16</td>
</tr>
<tr>
<td>services</td>
<td>3</td>
<td>0.4%</td>
<td>16</td>
</tr>
<tr>
<td>consultant</td>
<td>3</td>
<td>0.4%</td>
<td>16</td>
</tr>
<tr>
<td>company</td>
<td>3</td>
<td>0.4%</td>
<td>16</td>
</tr>
</tbody>
</table>
consulting
3
0.4 %
16

sydney
3
0.4 %
16

training
3
0.4 %
16

promotion
3
0.4 %
16

googgle
2
0.3 %
17

affordable
2
0.3 %
17

learn
2
0.3 %
17

specialists
2
0.3 %
17

matter
2
outsourcing
2
0.3 %
17

consultants
2
0.3 %
17

online
2
0.3 %
17

wikipedia
2
0.3 %
17

best
2
0.3 %
17

marketing
2
0.3 %
17

media
2
0.3 %
17

encyclopedia
2
0.3 %
17
<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>Percentage</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>free</td>
<td>2</td>
<td>0.3 %</td>
<td>17</td>
</tr>
<tr>
<td>gold</td>
<td>2</td>
<td>0.3 %</td>
<td>17</td>
</tr>
<tr>
<td>brisbane</td>
<td>2</td>
<td>0.3 %</td>
<td>17</td>
</tr>
<tr>
<td>help</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>leading</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>tools</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>webmaster</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>team</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>blink</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keyword</td>
<td>Count</td>
<td>Percent</td>
<td>Rank</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>linkfox</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>traffika</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>advice</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>service</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>provider</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>course</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>splash</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
<tr>
<td>works</td>
<td>1</td>
<td>0.1 %</td>
<td>18</td>
</tr>
</tbody>
</table>
optimization
1
0.1 %
18

global
1
0.1 %
18

solutions
1
0.1 %
18

roi
1
0.1 %
18

dejan
1
0.1 %
18

package
1
0.1 %
18

packages
1
0.1 %
18

put
1
0.1 %
18
2.5. December

BlogBook

dgm 1 0.1 % 18

sunshine 1 0.1 % 18

call 1 0.1 % 18

video 1 0.1 % 18

analysis 1 0.1 % 18

diy 1 0.1 % 18

specialist 1 0.1 % 18

portfolio 1 0.1 % 18

copywriter 1
BlogBook

0.1 %
18
ebooks
1
0.1 %
18
click
1
0.1 %
18
advantate
1
0.1 %
18
per
1
0.1 %
18
pay
1
0.1 %
18
divine
1
0.1 %
18
ramos
1
0.1 %
18
louie
1
0.1 %
18
40
©2013 dejanseo.com.au/blog
siterank
1
0.1 %
18

rfcd
1
0.1 %
18

friendly
1
0.1 %
18

organicwithseo|seo
1
0.1 %
18

page
1
0.1 %
18

anzsic
1
0.1 %
18

codes
1
0.1 %
18

arc
1
0.1 %
18
council
1
0.1 %
18

research
1
0.1 %
18

australian
1
0.1 %
18

your
1
0.1 %
18

firms
1
0.1 %
18

outsourcing
1
0.1 %
18

bruceclay
1
0.1 %
18

blog
1
0.1 %
18

works
1
2.5. December

0.1 %
18

melbourne
1
0.1 %
18

experts
1
0.1 %
18

data
1
0.1 %
18

advanet
1
0.1 %
18

pty
1
0.1 %
18

messageforce
1
0.1 %
18

processing
1
0.1 %
18

entry
1
0.1 %
18
outsource
1
0.1 %
18

rotapix
1
0.1 %
18

linxhurricane
1
0.1 %
18

solutions
1
0.1 %
18

consult
1
0.1 %
18

strategies
1
0.1 %
18

ltd
1
0.1 %
18

emarketing
1
0.1 %
18

around
1
0.1 %
18

net
1
0.1 %
18

news
1
0.1 %
18

get
1
0.1 %
18

latest
1
0.1 %
18

central
1
0.1 %
18

bpo
1
0.1 %
18

search
1
0.1 %
18

local
1
0.1 %
18

©2013 dejanseo.com.au/blog
tips
1
0.1 %
18

found
1
0.1 %
18

bizcatalyst
1
0.1 %
18

moves
1
0.1 %
18

goog|le
1
0.1 %
18

cbd|
1
0.1 %
18

high
1
0.1 %
18

visit
1
0.1 %
18
our
1
0.1 %
18

melbourne
1
0.1 %
18

value
1
0.1 %
18

link
1
0.1 %
18

agency
1
0.1 %
18

rankone
1
0.1 %
18

advertising
1
0.1 %
18

perth
1
0.1 %
18

building
1
Macro-Level Link Analysis

The next segment focuses on backlink analysis factoring in one of the strongest ranking factors, inbound links from diverse domain sources. Findings are quite interesting and show the players in the SEO industry with strongest backlink profiles. They are: ewebmarketing.com.au, searchenginerankings.com.au, roi.com.au, searchengineoptimisationworks.com.au and webprofits.com.au. These boys will be hard to catch up to unless Google changes something in their algorithm.
Among those with fewer links are dejanseo.com.au, freedomseo.com.au and quantumlinx.com.au which all have less than 1500 unique domain backlinks. Quantum Links have experienced some ups and downs but much like Freedom SEO they have had a sudden introduction into the first page of the rankings. Our guess-explanation for this (without going into a deeper analysis) would be rapid link growth with Quantum Links and overall marketing activity in case of Freedom SEO.

Language Complexity

Google has recently introduced a tool for simple website content complexity analysis. We won’t describe the exact method used in this article but it’s safe to say that the research is thorough enough and covers all of the indexed pages across all analysed websites.
As you can see only three companies feature advanced content: dejanseo.com.au, roi.com.au and searchengineoptimisationworks.com.au. This is, however, not necessarily a good thing if you are targeting small business. On the simple language scale seo.com.au, freedomseo.com.au and searchenginerankings.com.au are doing really well. This seems very fitting for one of them which focuses mainly on SMEs. Intermediate style language seems pretty even across all analysed websites with quantumlinx.com.au being at the top of the list and dejanseo.com.au and seo.com.au at the bottom.


### 2.5.2 Practical Guide to Viral Marketing (2010-12-20 23:27) - admin

[EMBED]

Giant Stars Viral Video by Dejan Petrovic & Brad Schwede

Practical Guide to Viral Marketing

Viral Content Development Example
Viral marketing. We’ve seen it, enjoyed it, passed it on, and most marketers have thought “I could do something like this for an xyz client.” Most of us, however, do not pass the “I could” stage while others experiment with varying degrees of success. The whole thing is a little bit chaotic and unpredictable. One thing is for sure, a well-executed viral marketing campaign can have a profound effect on traffic and branding. This post will reflect on main characteristics of successful campaigns and cover observations with two successful viral content experiments. Before jumping on the bandwagon you need to consider whether this strategy will be an effective use of your time and resources. Here are a few common scenarios worth exploring:

- Branding
- Positioning
- Product Launch
- Link Building

Although the effect of viral content can unleash incredible traffic, keep in mind that its targeting and immediacy in conversions and ROI isn’t likely to be as high in the initial stages. For example, if you are trying to create an immediate boost to sales of a specific product you’re probably better off running a “blitzkrieg” PPC campaign instead.

Selecting Your Content Type

For something to truly “catch on” you need to generate a truly original, brave and fresh concept. Here are some risky concepts:

- Controversy
- Scandals
- Sensationalism
- Conspiracy
- Taboo
- Dark Humour
- Shocking Material
- Sexual References
- Patriotism

Also on the relatively safe side we have:

- Comedy
- Games
- Tools, Services & Information
- Artistic Content
- Top Ten Lists
- Collections
Looking for Inspiration

There is no best way to do your research but to give your email address to “one of those” friends who constantly forward stuff to everyone else. Yes we are all sick of it and mostly delete emails titled “FWD: FWD: RE: Funny Cats”. Those things keep circulating, however, which means that they still have a high degree of popularity among internet users, workplaces, social networks and circles of friends. Pick the one (or more) you find could work for your industry and work on a new concept or... well simply look at ways or re-hashing the freestyle version into something you can control. Try not to kill the FWD spirit in the process by being sneaky. Stay open about your message and incorporate humour if applicable.

Media Format Choice

A good way to package things is to develop a flash movie or (if you have to) a PowerPoint presentation. Those media types (especially flash) tend to get less edited in the process which ensures the integrity of your message is untouched and prevent any negative distortions. People will of course do it if they really want to, but if you make it too easy it’s almost an invitation to get creative at the cost of your brand. If you are creating a video, maybe incorporate a non-intrusive logo or watermark throughout the length of the feature followed by the end screen credits. If you have a flash file link it to a permanent URL (which can of course be redirected) at the end of the feature. Multimedia content will be plucked and hosted elsewhere if it’s interested enough and you don’t want to lose its primary function in that case.

Portability Issues

Decide whether you want the content to be freely distributed or locked to a domain. If the sole purpose of the viral is to be a link bait then you would certainly want to lock it to a domain name (unless we’re talking about a site/blog widget that contains your link). Otherwise if we’re talking about increasing brand awareness then by all means let it spread everywhere and use tools like “Send-to-Friend”, social media buttons, feeds and copy and paste code snippets for bloggers and site owners.

Be Ready

If you spend lots of money and time on development of your viral, make sure that your technology doesn’t let you down when it takes off. Make sure your hosting arrangements allow for scalable bandwidth. A popular video or a flash file can easily consume a terabyte of bandwidth in a short period of time.

Deployment & Expectations

Probably one of the best ways to launch a viral is to send it to your friends and colleagues, sit back and watch. In case things aren’t going as you expected, give it a light nudge with your site(s) and social media. At the end of the day if it doesn’t take off at all, do a post-mortem and learn for the next time. If the content proves to be good enough the viral will soar and traffic will peak. This usually happens when the link appears on a big traffic news website or a front page of a big social bookmarking website. The initial traffic will surprise you and then likely exceed your expectations, after that, there is an eminent drop followed by a steady flow of traffic from a vast variety of links accumulated in this short period of time.

Viral Experiment I: Acid Trip

I conducted [1]my first viral test a few years back after receiving an email linking to a page talking about an old government experiment with an artist under influence of LSD. The original set of drawings was packaged up in a single flash file with the usual back and next buttons, selectable flash text and a link to a target page on the final click. The flash movie was simply linked from the main site’s home page. No other means of promotion were applied. This simple flash movie gradually became a hit among savvy users and earned lots of natural links which pushed the page up in the search engines (search term: Acid Trip) which in turn
resulted with even more traffic exposing it to a wider range of audiences. I eventually decided to make use of this traffic and promote my favourite online radio station.

Viral Experiment II: Giant Stars

The second test has been conducted deliberately with a great deal of research and planning. For me it was important to know whether it would be possible repeat the success of the first flash, not by accident but on purpose. This piece uses a superlative approach which usually works well (eg: biggest, tallest, fastest, ugliest...etc). There was a twist to it though, it also contained the “wow” factor – a comparative element that blows people away as they go through it. This is what I believe was the key component that contributed towards the success of this experiment – a comparison of things and play on audience’s perspective. So, to get to the point – I used an already circulating set of images comparing the sizes of planets of the solar system, their size relative to the sun, and likewise, our sun measured up against some pretty big celestial bodies out there. This time the page was promptly distributed via usual social media channels and eventually ended up on the front page of one of the main sections of StumbleUpon. This brought a wonderful flow of traffic to the site and within a few days generated lovely natural links. Soon after that we rolled out a video version which was distributed via TubeMogul. I monitored the traffic growth and decline with great interest and attention and am amazed that once stabilised, the traffic continues to flow steadily for years.

Conclusion

Bottom line - it’s all about the quality of your content. There are things that you can do to make the whole process smoother though, so plan your viral carefully and make sure you’re not harming your brand in the process. There are lots of little annoying things that can get in a way, you will spend time and money to develop it, so do take time to also optimise your viral. Finally, sit back, relax and watch what happens next!

Taking Action

Viral marketing is an extraordinary internet marketing strategy for traffic generation, link building (SEO) and brand building. If you believe your business could benefit from a viral marketing campaign contact us on 1300 123 736 and we will explore your ideas further.

Related Links:


2.5.3 The End of Google PageRank (2010-12-31 18:15) - admin

The first version of the PageRank technology was created while Larry and Sergey attended Stanford University, which owns a patent to PageRank. The PageRank patent expires in 2017. We hold a perpetual license to this patent. In October 2003, we extended our exclusivity period to this patent through 2011, at which point our license will become non-exclusive. Despite many claims by SEO companies around the world that PageRank does not “work” anymore, we see Google employees mentioning it over and over again in articles, blogs and videos. Webmasters have been wondering about Google’s intentions around PageRank due to the fact that the last public update was back in April 2010. The usual trend of public updates was around four times per year. According to [1]this document Google PageRank patent is in fact owned by [2]Stanford University and it expires in 2017. In October 2003 the patent exclusivity was extended to 2011 and after that the license will become non-exclusive. How this affects the search engine algorithm and the ranking signals is yet to be seen, but many speculate that this is...
the end of PageRank in the shape and form we know today.

Update:

We like a comment made by Chris from [3]http://www.websitepublisher.net/made on an SEO Roundtable [4]post similar to this one. Chris says: "I don’t think Google’s ranking algorithm, at this point, is very relevant to their success or their position. Their main advantage, at this point, is their infrastructure. Brin and Page did not just develop a good ranking algorithm, but they also developed storage mechanisms for data recall, which is a significant portion of running a search engine. Supposing you gave some startup Google’s exact algorithm, they wouldn’t be able to use it because they lack the processing power Google has. That is a very big deal.”

Chapter 3

2011

3.1 February

3.1.1 Influencing User Behaviour through Search Engine Optimisation Techniques

Author: Dejan Petrovic
Published on: 2 February 2011

1 - Introduction

“You’ll sometimes want to attract robots, or lace your website with spider bait.” (Lieb, 2009)

Grappone and Couzin (2011) define search engine optimisation (SEO) as a diverse set of activities one can perform to increase the number of desirable visitors to visit your website via search engines. Just as all people who open up a shop or any business are interested in attracting clients above all else, all developers of websites, be it personal or business websites are interested most in attracting visitors to their site. Most businesses and web developers have come to understand that the best way to attract customers it to know how consumers behave and find ways to influence them. An important aspect of online user behaviour that researchers have noted is [1]how consumers search for content online. 85% of consumers, according to Vertexera, go to the most common search engines (Google, Yahoo and Ask!) to look for vendors or suppliers of a good or service. The user then usually looks for the results given by the search engine, with 73% of consumers never going beyond the first page of the results given by the search engine, with many of these only looking at the top half of the search engine webpage (Petrovic, 2010). The main purpose for Search Engine Optimisation is to make sure a website’s product or information appears as high as possible when a search is performed on a search engine using as many keywords as possible. The main aim of this paper, therefore, is to establish the styles and benefits of using SEO to positively influence user behaviour.

1.1 - Search engines

Before launching into how user behaviour is influenced by SEO, it is important to first understand how search engines work. There are two types of search engines: crawler-based and human-powered search engines. Crawler-based search engines sends ‘crawlers’ to search through websites. Crawlers, at times called spiders or robots, are programs or automated script that scuttle through websites finding and copying text in them to create a vast index in a process known as ‘spidering’ (Lieb, 2009). This index is essentially a database from which a search engine will use to create results as fast as possible. Google uses this technique to create its
search engine results. Human-powered search engines, on the other hand, require a short description that will be placed on the search engine directory. This description is usually written either by the company running the website or by reviewers. The most famous directory that exits as such is LookSmart. Some hybrid search engines such as Yahoo and MSN provide both crawler-based and human-powered results.

1.2 - Online Consumer Behaviour

It is also important to understand online consumer benefit as it will act as a guiding theme throughout this research paper. A study conducted by Cotte, Chowdhury, Ratenshwar & Ricci in 2006 revealed that among the visitors of websites, there are four distinct intention and motivation categories.

- Exploration.
- Information.
- Entertainment.
- Shopping.

Petrovic’s research shows how there are three main steps that users who go online take when looking for information about a brand online.

1.2.1 - Step 1: Initial search

Consumers’ first search for certain keywords associated with a product indicated that a significant percentage looked for product brands. It is by search for the available brands of certain products that consumers start looking for a product. This search is always done online starting with the most common search engines like Google, Yahoo and Ask!

1.2.2 - Step 2: Finding the product

After successfully search for the brand, the consumers then sift through the available search results, with the highest percentage of potential customers visiting the top results of the search results. Only 7% of the customers bother to ever to beyond the third page of the search results.

1.2.3 - Step 3: Practical testing

Once online users click a website of interest to access it, they can then use the available data in the website to find the necessary information they need about a product.

However, there is an issue that Petrovic observes among online consumers: patience. Most online users are increasingly becoming less patient in waiting for a webpage to load. According to his research, most users are only willing to wait 5 seconds for the webpage to load. The following pie chart shows how long consumers are willing to wait for web pages to load.
2 - Benefits

2.1 - Branding

One way of attracting customers and therefore optimising the number of clients that visit a website is building a strong online brand. Rob Frankel, described as the top branding expert on the planet, describes branding as not about getting your prospect to choose you over your competition; but about getting your prospect to see you as the only solution (Gerhart, 2002). A high ranking in a search engine means there is a higher chance that clients get to see the product a company is selling even if they do not buy it. In fact, twice as many consumers are likely to recognise a business ranked in the top 3 of a search engine result than those appearing in prominent banner space (Vertexera Inc.). Not only is online marketing a lot less expensive than offline marketing, it also exposes the product a company is selling to a huge market.

By using techniques like predictive heatmaps and keyword manipulation, a brand is bound to appear at the top of the search engine results. The main benefits of this are:

1. When a brand appears near the top of search engine results, more people are bound to see it and therefore build an association between the brand and certain needs. For example, if clients are looking places to buy ceramic cups online, the keywords they are likely to search are ‘cups’, ‘ceramic’ and possibly ‘home delivery’ or ‘cheapest’. By optimising their website to have these keywords captured by search engines and have their brand appear first on the search engine, companies will have internet users who search frequently for these products online know the following about their brand.
   
   (a) The company has best ceramic cups in the market.
   (b) These ceramic cups are most likely relatively well priced compared to other cups in the market.
   (c) The company offers extra services like home delivery and any other service the company may have.

2. Once clients are aware of these facts, they are most likely to click the company’s website when searching for this product and as a result are more likely to visit the website. It is important for websites to also improve the ease with which customers can navigate their websites as this improves the relationship...
customers will have with the company’s website. Having websites that are full of graphics and are therefore hard to navigate, or having websites that do not have an adequate number of inbound links would quickly discourage clients from ever visiting the website (King, 2008).

2.2 - Direct target

Another benefit of SEO is increased direct target traffic. By optimising a website by using certain keywords, the website targets certain customers only by ‘weeding out’ online users not interested in the product. A website selling sports shoes only, for example, will have keywords like ‘sport shoes’, therefore getting rid of untargeted clients like those looking for leather shoes, sandals, sneakers, loafers etc.

2.3 - High Return on Investment

SEO services are often provided by companies that help websites optimise their websites for business. Unlike using banner ads, research shows that SEO costs a lot less yet increase the number of clients almost exponentially (Enge, Spencer, Fishkin, & Stricchiola, 2009). Research done by Vertexera Inc. has shown that banner ads cost between $2,500 and $35,000 while SEO costs about $1,000. However, as discussed in the previous section on branding, clients are more likely to notice results of a company on a search engine than banner ads. A gastric bypass surgeon who optimised his website for search engines had the number of clients he generated from his website increase from 2-3 per week to 50-70.

2.4 - Long term visibility

Banner ads are usually placed on websites and online advertising spaces for a certain period of time, with that period paid for by the company. Once that time is done, the advertisement is pulled down unless the company pays more to have it kept up. This means therefore that if a company depletes its marketing budget, the ad will no longer be available for clients to see. On the other hand, if a company maintains good SEO practices and makes sure its website’s content remains relevant and up-to-date, the chances of having it appearing among the top 5 results for a search on any search engine could remain high for as long as possible.

3 - Combining usability and SEO

3.1 - Website usability

Usability refers to the way one uses or experiences a website (Clay & Esparza, 2009). A good website is supposed to be usable to a large part of the online population, with clear links and a well-done layout. The following are ways in which a website can be tweaked to improve its usability.

- Simple navigation system – the navigation system of a website has to be simple to navigate, with appropriate text links, consistent menus and a search box.

- Clear content – content of a website should be clear to read or even just scan. There is a reason accessing the internet is called ‘browsing’. Most internet users hardly spend time reading content of a website. Content should be kept minimal and colour contrast should be kept neutral to avoid tiring the visitors of the website easily.

- Support the brand – it is important to make the website about the brand, not the company/individual. A website is supposed to promote brand purchase intent since it is the purchase intent that is supposed to link the brand and the response (Jaffe, 2003).

- Provide client feedback channels – this is probably the most important step. Clients should be able to provide the company with feedback regarding the website since they are the ultimate deciders on the
success of the site. It is important for clients to advise the company how they find the websites ability to aid with decision making, its shopping cart limitations and abilities and how it stimulates them to buy the product. Testing the site on real people before launching it could help improve these features of the website a great deal.

3.2 - Combining with SEO

Up until recently, most web developers and users never saw the link between search engine optimisation and usability. The goal of building any website is making sure it is simple to use for all users and it gets as much traffic as possible. Many scholars, however, argue about which is more important, improving usability or improving SEO. The truth is that a balance of both works best. In fact, if cleverly done, improving usability has the potential to go a long way in improving SEO. The first step is to find out what SEO and usability have in common then put them into balance. The key to creating a useful SEO is not just having keywords showing in your website but having keywords that are relevant to the content of the website.

The table below shows how improving website usability often improves the SEO of the website as well.

<table>
<thead>
<tr>
<th>Usability improvement</th>
<th>SEO benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researching to find out where clients are looking for the company.</td>
<td>Combines with keyword search.</td>
</tr>
<tr>
<td></td>
<td>Building a network of links coming from external web pages to increase the number of people that can find the company’s website</td>
</tr>
<tr>
<td></td>
<td>Gives search engines the perception that the website is an expert website therefore increasing link equity.</td>
</tr>
</tbody>
</table>
Improving the usability of the site.

Helps search engine spiders find keywords of the site easily.

Writing clear pages on what the site is offering

Helps search engines determine what the site is about

Uploading the site to fast reliable servers to provide good site performance to users

Improves the speed at which search engine spiders access the website (Quick, B. 2011).

4 - Conclusion

The aim of this paper was to determine how user behaviour online can be influenced by improving a website’s SEO. This paper targets companies who want to improve the number of products they are able to sell online by improving the ranking their websites receive on popular search engines such as Google, Yahoo, MSN and Ask!

The paper starts by explaining important themes upon which the paper would be constructed. It then goes to explain the benefits of SEO and finally how usability of a website can be combined with SEO to bring out a website that best benefits the business.

Works Cited


3.2 March


PDF Version:


3.2 March

3.2.1 Obstacles in Experimental Testing and Reverse Engineering of Google Algorithm (2011-03-22 13:39) - admin

One tricky thing about SEO as a marketing discipline is that it’s not an exact science and we’re dealing with proprietary algorithms which grow more complex each year. Despite the strong community in the industry and contributions to common knowledge, SEO professionals struggle to see the whole picture and are often puzzled by sometimes strange behaviour of the biggest search engines in the world.

In this article we will outline and comment on results from an experiment Dejan SEO team performed in 2008. Before we get into data let’s consider the following factors which (among others) drive advancement of search engines today:

- Spam
- User technology adoption
- Diversification of online channels
- Growing amount of online content
Spam

Each major wave of search engine incarnation is followed by a tsunami of spam forcing search engines to stay agile and combat manipulation. Spam has been and will be one of the major driving factors in advancement of search engines.

User Technology Adoption

Search engines have to keep up with growth and diversification of user types and platforms forcing them to adapt to different languages, geographic locations, browsing platforms and temporal factors.

Diversification of Online Channels

Blogs, forums, local directories, job websites, social networks... search engines are having so much fun feeding on this data and so much headache in trying to make out what’s what, validate, cross reference and use as a signal. This will never change and technology will continuously advance – be prepared for continuous adoption and adaptation.

Growing amount of online content

Google is the first search engine that put true emphasis on the sorting of results and that was a good thing as this became useful with a growing number of online resources which needed to be prioritised, firstly in order of relevance to users’ search query and secondly by page’s importance (e.g. PageRank, also referred to as ‘PR’).

What is the problem for search engines face?

The main problem search engines face today is rapid influx of new data and growing number of online resources. ‘Prosumerism’ is taking momentum and everyone has something to say – each networked individual is a micro content publisher. Computational resources and bandwidth are still limited so there is always the matter of prioritisation and optimisation of resource usage. Algorithmic changes are essential in order to set crawling, indexation and result serving priorities right. The above mentioned factors are keeping search engines busy and add to complexity of our job – which is to understand search engines and enable better exposure for our or our clients’ websites.

[1]

The Experiment

In order to reveal a small piece of the puzzle we decided to test Google’s crawling behaviour by monitoring the flow of PR through a complex iterative navigational structure. A new domain was registered for the purposes of the experiment and fed with a single PR7 link pointed directly to the index page. No other links were added in order to preserve the integrity of the experiment. The estimated flow of PageRank is illustrated below:
This estimate is based on PR flow observation of a sample of hundred random domains. Typically toolbar PageRank (TBPR) value reduces by around 1 point for each level deeper from the highest source of PR and where the highest source of PR is the home page. Actual PR value is estimated to be in average 25% higher or lower than the visible value. Actual flow of PageRank seems surprisingly different. Pages marked in grey are those where TBPR value was as expected, green pages have higher value and red lower value than expected.
The surprise first comes in the third layer where seven pages get PR5 instead of PR4, perhaps due to the fact that the initial source of PR had value closer towards 7.5 and the fact that the site has only four second level pages to share internal PageRank with. Observe that links closer to "Home" in the navigation (or left side of the horizontal menu) seem to be getting more PR passed to them. Due to the fact that TBPR values are rounded to a single digit this is only evident by observing its distribution in the third level of cascading site architecture. Observe the last node of the last page of the top level navigation (the only red item in the third layer). At this point is where Google stopped assigning PageRank to pages, except for the last page in the fifth layer which for some strange reason scored a PR3, hinting at a possible external link affecting the experiment. To discover why Google randomly stopped assigning PageRank to pages our team went to observe the caching behaviour of the site throughout the period of 3 months.

Week 1:

In the first week since PR7 link is placed on the site only index page is accepted in Google cache, hinting at potential balancing of resource usage on Google’s behalf for this newly discovered resource. Given that a high PR source linked to this site (which led to its discovery) we expected more rapid caching of all pages and this cautious behaviour came as a surprise.
Week 2:

By the end of the second week new pages are visible in Google cache. Not surprisingly the entire first level is in cache. What is interesting that only two pages of the third level (2.1 and 3.1) enter cache, the rest remain un-cached. At this point we’re starting to make the connection with subtle PageRank distribution variations and its effect on the caching rate.
In week four is when the strange caching behaviour becomes clear as the rate of caching now clearly explains the higher than expected PageRank value for the page 5.4.3.2. What remains unexplained at this stage is why this page got this PageRank in the first place and we’re looking for potential external links that may have affected the experiment - so far nothing found.
3.2. March

Figure 5: Indexation Rate in Week 4

Week 8:

It’s finally in the eight week of the experiment that we’re seeing full site in Google cache to its deepest architectural level. At this stage PageRank has updated and revealed it’s unusual allocation as see in the Figure 2.
Conclusion

This experiment demonstrates how Google approaches new sites and allocates its resources according to the level of trust for newly discovered resources. In this simplified model we’re not observing many other factors which can influence search engine behaviour and focusing on cascading behaviour of Google PageRank and its effect of resource allocation and the rate of caching of discovered content. The apparently random element in content indexation and assigning of PageRank which in addition to variable value of PageRank causes indexation to suddenly stop and resumes at later stage (perhaps when crawlers internal limit has been reached) appears to be a complex set of variety of internal rules rather than a purely random behaviour. Some degree of random behaviour could potentially be beneficial to a search engine as it would provide a platform for organic growth of the algorithm much like sporadic mutations in gene replication enable living species to evolve. Secondary benefit could be a layer of protection against deliberate reverse-engineering of its algorithm as each attempt to probe the system would result in subtly distorted version that would not match previous attempts. To prove this more experimentation is needed, perhaps the same experiment repeated more than one time to observe differences in behaviour under same circumstances.

References


Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999) The PageRank Citation
3.3 May

3.3.1 Domain Hyphenation Test (2011-05-02 01:05) - admin

There has been much talk about domain hyphenation with Google’s response vague as usual. Dejan SEO team put this to test and in conclusion non-hyphenated domains have an advantage over hyphenated ones. For test specifics and results read on.

[styledbox type="general" ]

Update

Results of this test were published on SEOMoz on the 1 May 2011. The status of the ranking domain has changed since then in favour of the hyphenated domain. The advantage of this domain was the first position on the page. Since our test domains did not have their own title tag they inherited anchor text as their title. Notice the discrepancy between the title and the domain (love/loves) which does not seem to represent a problem and Google highlights both as a match in SERPs.

Domains Used


Search Results


Linking Page

- Anchor text: "Dejan SEO love testing"
Here are the technical details of the setup:

Domain Registration & Hosting

We registered two domains:

- keyword1keyword2keyword3keyword4.com
- keyword1-keyword2-keyword3-keyword4.com

They were registered seconds apart and in order as above using same whois information and custom name servers:

- ns1.keyword1keyword2keyword3keyword4.com
- ns2.keyword1keyword2keyword3keyword4.com

and:

- ns1.keyword1-keyword2-keyword3-keyword4.com
- ns2.keyword1-keyword2-keyword3-keyword4.com

Both domains were hosted on a different IP of the same C-block.

Content

Both index pages contained only one line in the following format:

Did you know that keyword1keyword2keyword3keyword4.com
Did you know that keyword1-keyword2-keyword3-keyword4.com
No HTML.

Links

Both test domains were linked from the same page on another domain. In the source code of the linking page, the first link was the hyphenated one, giving it the initial (if any) advantage over the other.

Early Results

- Day 1: Initial search brings up only dailychanges.com who picked up the domains.
- Day 2: No change.
- Day 3: Hyphenated domain is in the index, can be found below dailychanges.com
- Day 4: Hyphenated domain is in cache, under dailychanges.com still
- Day 5: Non-hyphenated domain enters Google index (found through alert), cached and ranking #1
- After 3 months: Hyphenated domain gains position #1

Possible Undesired Influences

- Non-linked references dailychanges.com
- Nofollow references from dailychanges.com
- Order of links in HTML of the linking page
- Order of domain registration (seconds apart)
- Discovery by Googlebot
- Similarity of content

None of the above seem to have affected the test.

2. [http://dejan-seo-loves-testing.com/](http://dejan-seo-loves-testing.com/)

### 3.3.2 Impact of Rankings on SERP CTR (2011-05-15 14:14) - admin

There has been very little original research in the area of position-based clickthrough rates (CTR) in search engine result pages (SERP) and most SEO companies still quote outdated information ([http://www.seoresearcher.com/distribution-of-clicks-on-googles-serps-and-eye-tracking-analysis.htm](http://www.seoresearcher.com/distribution-of-clicks-on-googles-serps-and-eye-tracking-analysis.htm)) which may not even be relevant anymore. Dejan SEO team has performed new research in this field in May 2011 and announced the latest findings to the rest of the SEO community. Summary of our research is represented in the SERP CTR heatmap below:
Note: The above figures are separate averages for a range of 50-300 selected phrases for each position and the total does not necessarily need to add up to a 100% value. There are numerous factors which impact CTR values but this goes beyond the scope of our study.

Technical Data:

- Search Queries: 1,000,000+
- Ranking Pages: 200 per position
- Position Type: Ranking Average
- Industry: e-Commerce / Retail
- Search Sources: Google (Worldwide)
3.3. May

- Domains: Australian (.com.au)
- Tools: Google Webmaster Tools, Custom Scripts

Related Research


CTR Data (Table Format):

<table>
<thead>
<tr>
<th>#</th>
<th>CTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43.2%</td>
</tr>
<tr>
<td>2</td>
<td>30.7%</td>
</tr>
<tr>
<td>3</td>
<td>23.3%</td>
</tr>
<tr>
<td>4</td>
<td>19.7%</td>
</tr>
<tr>
<td>5</td>
<td>15.1%</td>
</tr>
<tr>
<td>6</td>
<td>14.3%</td>
</tr>
<tr>
<td>7</td>
<td>11.4%</td>
</tr>
<tr>
<td>8</td>
<td>10.1%</td>
</tr>
<tr>
<td>9</td>
<td>8.9%</td>
</tr>
</tbody>
</table>
3.4 June

3.4.1 schema.org (2011-06-03 20:18) - admin

Google, Bing and Yahoo have come up with a new initiative in order to assist with a common set of schemas for structured data markup. This resource is designed for webmasters to enable them to easier present their information to search engines.

Google already supports structured markup and examples can be seen in via rich snippets which pop up in reviews of products, events, recipes, movies...etc.

Supported markup types include:

- Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries
- Embedded non-text objects: AudioObject, ImageObject, VideoObject
- Event
- Organization
- Person
- Place, LocalBusiness, Restaurant
- Product, Offer, AggregateOffer
- Review, AggregateRating

Full list is available here: [1]http://schema.org/docs/full.html

Usual treatment of microdata, microformats and RDFa has changed and with schema.org it has been decided in favour of only microdata as main standard. This should not cause concern with those who already have other standards implemented as Google promises support for those as well. Do not mix different standards on the same page though.

To test out your rich snippets use this tool:


1. [http://schema.org/docs/full.html](http://schema.org/docs/full.html)
Two years ago Grzegorz Czajkowski from Google’s system infrastructure team published an article which didn’t get much attention in the SEO community at the time. It was titled "[1]Large-scale graph computing at Google" [1] and gave an excellent insight into the future of Google’s search. Here I will highlight some of the little known facts which lead to transformation of Google’s algorithm in the last two years.

Everything is a Graph

Czajkowski argues that much of our existence can be represented in a graph form including social activities, personal relationship, professional activities, technology, transactions and other aspects of our lives, much like citations work among scientific papers. Google was the first company to successfully harness the power of the web by applying advanced graph analysis techniques including the concept of PageRank.

Here’s an interesting fact. The concept of mapping society mathematically is nothing new. There are countless theoretical and even fictional references to calculations of socio-economic and behavioural elements in attempt to understand the humanity and map the future path of our civilisation. One such concept comes from the famous science fiction writer and a scientist Isaac Asimov in his Foundation series. Asimov’s name for what Google’s heading towards is "psychohistory".

"Psychohistory is a fictional science in Isaac Asimov’s Foundation universe which combines history, sociology, and mathematical statistics to make general predictions about the future behavior of very large groups of people, such as the Galactic Empire. It was first introduced in the five short stories (1942–1944) which would later be collected as the 1951 novel Foundation.” Source: http://en.wikipedia.org/wiki/Psychohistory_(fictional)

In Asimov’s fiction psychohistory is mapped towards a "prime radiant” which may deviate due to many variables and the scientists occasionally adjust the formula to factor in new variables.
Dealing With Complexity

Static graphs in science can be extremely complex but when it comes to a dynamic environment such as society, knowledge, communication and information graphs exhibit perpetual growth in size and data/relationships which introduce a whole new dimension to the problem of scalability and computational power.

Google’s solution to scalable graph analysis takes inspiration from the Bulk Synchronous Parallel Model in parallel processing. The framework is called Pregel and is capable of mining a wide range of graphs through a unique iterative vertex treatment. In Pregel each graph vertex works independently and can receive and send messages from and to other vertices at different stages of iteration which enables self-modification and mutation of graph’s topology without having to re-run the entire process from scratch. With this framework Google is capable of incomparable computing scalability when it comes to graph data and it simplifies calculation of PageRank. Below is the code sample from the official Pregel paper [2]:

```java
class PageRankVertex {
    public void vertex(String vertexId) {
        float PageRank = 1.0f;
        int maxIterations = 100;
        int iterations = 0;
        int minIterations = 2;
        double dampingFactor = 0.85;
        double alpha = 0.15;

        while (iterations < maxIterations) {
            double oldPageRank = PageRank;
            PageRank = dampingFactor / numVertices;
            for (String neighbor : neighbors) {
                PageRank += 0.15 / (float) numVertices;
            }
            if (iterations % minIterations == 0) {
                previousPageRank = PageRank;
            }
            iterations++;
        }
    }
}
```

Figure 4: PageRank implemented in Pregel.

[2]Michael Nielsen [3] has recently covered Pregel and elaborated on its practical use (calculation of PageRank) and provides useful pseudo-code samples. According to Nielsen, one of the main benefits of the new framework is the fact that it removes the need for manual intervention from programmers as it is capable of scaling on a cluster autonomously. This leaves software engineers with more time to focus on the algorithm itself. This proves that Google hasn’t changed their view on human element in their software and offers strong hint that even social factors will always remain just maths and nothing more.

Computing PageRank

[3]PageRank [4] is based on the [4]random surfer model [5] where it is assumed that a person browsing the web follows links in a linear fashion until the point where their interest stops to the point where they stop browsing or abandon the current tree of research and start elsewhere (back to search results). With this logic (non-native) PageRank reduces with each click away from the source document. Naturally this is a simplified example in which there are no external links pointing to any of the documents along the browsing path. So for example in a website with a home page PageRank of 2 the next click may lead to a PageRank 1 page and finally to a zero value page at which point the “interest” to the user becomes minimal. Typical value of the PageRank dampening is 0.15. This tends to be more complex in a typical web document as there is usually more than one link on the page and PageRank tends to circulate throughout the sites’ navigation (sitewide links).
In Pregel aided PageRank calculation main data input is the web graph generated through crawler based activities. In practical terms, one can treat the web as "the graph", all of its documents (pages and indexable files) as "vertices" and links as "edges". In this model vertices (web documents) are typically initialised with a starting value. The interesting point is that this initial value makes no significant impact on the end-result in the computational model. After initialisation Pregel runs through supersteps by updating own value and sending messages to other vertices.

Nielsen also points out that each vertex also has a halting property, a value which determines the status as active or inactive (at which point the computation stops).

Related Frameworks & Methodologies
Bill Slawski from [5]SEO by the Sea commented on this article and submitted a presentation by David Konerding from Google titled: “[6]Think Like a {Vertex, Column, Parallel Collection }”[8]. This document proves that there is much more behind Google and in addition to Pregel (the large-scale graph processing framework) they employ interactive analysis of web-scale datasets through Dremel and utilise FlumeJava for data-parallel pipelines.

Due to limitations in flexibility and application of MapReduce [7], software engineers had to design their own toolkits and frameworks for data-intensive parallel processing, typically with multi-step graph operations on a large scale and complex flow datapipe. It was interesting to find out that they also had to employ own tools to assist data analysts who deal with enormously complex (trillion-row) datasets.

The Pregel part of the presentation gives a good overview and introduces its similarity with [7]Parallel Boost Graph [6]. From what we understand Google uses Pregel widely within their systems as it’s easy to program and ‘expressive’.

Benefits:
• [8] Breadth-first search

• Strongly connected components

• PageRank Compatibility

• [9] Label propagation algorithms

• [10] Minimum spanning tree


• Several kinds of [13] vertex clustering

• [14] Maximum and maximal weight bipartite matching

Dremel: Think Like a Column

There is no better way to explain the concept other than by illustration. In the 3D renderings below you can see the logic behind the record-oriented and column-oriented approach:

In the next example recordsets are arranged in distinct columns:
Characteristics and Benefits of Dremel:

- Trillion-record, multi-terabyte datasets
- Scales to thousands of nodes
- Interactive speed
- Nested data
- Columnar storage and processing
- In situ data access (e.g., GFS, Bigtable)
- Aggregation tree architecture
- Interoperability with Google’s data management tools (e.g., MapReduce)

Practical Application of Dremel at Google:

- Analysis of crawled web documents
- Tracking install data for applications on Android Market
• Crash reporting for Google products
• OCR results from Google Books
• Spam analysis
• Debugging of map tiles on Google Maps
• Tablet migrations in managed Bigtable instances
• Results of tests run on Google’s distributed build system
• Disk I/O statistics for hundreds of thousands of disks
• Resource monitoring for jobs run in Google’s data centers
• Symbols and dependencies in Google’s codebase

FlumeJava: Think Like a Parallel Collection

FlumeJava was released to Google users in May 2009 and now hundreds of pipelines who process gigabytes to petabytes are run by hundreds of users every month. It is easier to use than MapReduce and can control optimizer and executor when needed (as well as better handling of unpredicted situations). Example below is for TopWords:

```
readTextFile("/gfs/corpus/*.*.txt")
.parallelDo(new ExtractWordsFn())
.count()
.top(new OrderCountsFn(), 1000)
.parallelDo(new FormatCountFn())
.writeToTextFile("cnts.txt");
.FlumeJava.run();
```

Qualities Google is after are fault tolerance by design and handling the processes in such way that if individual node fails it merely slows down completion (instead of stopping it altogether). Everything is large-scale by design and architecture (trillions of rows, billions of vertices, petabytes of data). Tools and systems within Google are interchangeable and used by multiple groups.

Conclusion

There is no doubt that this is only a glimpse at Google’s true complexity in their efforts to organise the world’s information. Frameworks of this kind enable Google to maximise the use of their resources and provide superior results in comparison to their competitors, not just due to crude computing and processing power but also thanks to intelligent software solutions. Google is still not abandoning search as their first priority and their next obvious target with this great framework is to understand social interactions online - and in the way that nobody else can.
For Google, all the stars are aligned, they have the technology, money and all the data in the world. Soon their technology will likely reach the level incomprehensible to us ordinary humans. Credit goes to two technological visionaries Ray Kurzweil (Technological Singularity) and Isaac Asimov (Psychohistory) who predicted it all years ahead of us all, and I shall end with my favorite quote:

"Any sufficiently advanced technology is indistinguishable from magic.", Arthur C. Clarke

References


[8] Konerding, D., - Think Like a {Vertex, Column, Parallel Collection }

3.5 August

### 3.5.1 Search Algorithm Release Process at Google (2011-08-06 12:31) - admin

During the panel discussion at the Churchill Club Danny Sullivan asked Matt Cutts, Ben Gomes, and Amit Singhal about the recent algorithm changes including the famous Panda. Although the answers were not particularly informative they did give insight into Google’s quality algorithm release process.

According to Amit Singhal, once identified the problem is solved strictly by algorithm (Google still claims they do not use manual human intervention for much more than testing the quality of results and manual penalties in very few instances). Candidate algorithms are then implemented in code and tested using old query logs. There results are then analysed for quality. In the next stage small human control group is used on sandbox results which show A-B in random fashion. The data collected is used towards a limited live release where a relatively small number of users may see new algorithm in action. These results are further analysed and assessed for quality before making the algorithm permanent. It’s likely that these mini releases happen often and may be responsible for sometimes surprising search results many SEO professionals are seeing in SERPs.

Full video:


1. http://www.youtube.com/embed/J5RZOU6vK4Q?rel=0
3.6 October

3.6.1 The Future of Search: Interfaces (2011-10-17 21:40) - admin

According to Vernor Vinge and Ray Kurzweil [1] we’re at the dawn of technological singularity. According to their theory, in a few decades (circa 2045) technological advancement will reach a critical mass and result in an explosion of advancements in fields of science and technology.

The Faith of Search

Rather than beating around the bush, I’ll get straight to the point. The ‘search’ will eventually die or rather evolve into ‘knowing’. What knowing exactly means is hard to say, but we can speculate around the topic. Let’s say that it might be a combination of real-time data stream validated and cross-referenced against rudimentary axioms, trending theories, current affairs and network of trusted ‘validators’ in official and social capacity. The truth has always been a flexible and often personal perception. In a society where we’re all connected many truths will be amalgamated and normalised by the power of majority and top-level ‘influencers’ including various authorities on different subjects.

Fossils of Knowledge Acquisition

Fossil remains often reveal strangely formed ancestors of today’s animals. Some of them are now extinct but some have proven to be a success. Same thing will happen with knowledge acquisition methods, search and various interfaces which will aid in the process.

We’re way past painfully slow connection speeds of dial-up modems and entering an era of mobile computing, portable and wearable electronics and communication devices. Information is handy and quick to reach. What is getting in the way of progress is clumsy interfaces. You either have the keyboard or not. Without it you’re stabbing clumsily at the touch screen, dictating to a faulty voice recognition system or retro-writing / swyping on your screen. With keyboard you’re limiting your portability.

We really need to find a way to get rid of hardware, as much of it as possible - or at least minituarise it to the point where it does not get in a way anymore.

Evolution of New Interfaces

While studying at Griffith University I formed a group dedicated to research in EEG systems. Electroencephalography is not exactly a fully applied science but rudimentary technology existed since late 20th century. Hooking up a few electrodes in a shielded room, recording brain activity and feeding the data into our infant-like neural network. Except, we did not get clearance for human subject experimentation. The project died and so did my interest in thought-control interfaces. I was waiting for somebody else to continue development, and many around the world did (with partial success, clumsy equipment and enormous running and maintenance costs).

Brain Control for the Masses

Significant attempts at brain cortex activity visualisation have taken place over the last few years [2]. This year, a group of scientist have published their research on visualisation of brain activity [3] and uploaded output to YouTube. The results are stunning, even hard to believe. Although we’re far from being able to record the dreams, this experiment shows that significant advancements have taken place in the EEG
technology in the last decade.

This is nothing in comparison with the ever-increasing number of commercially available (and affordable!) brain interfaces such as EPOC [4] by Emotiv which comes with an SDKs and even an API access for the developer community. This product is already available in the United States and pending approval for Australia and other target countries.

Watch the video:

IFRAME: [1]http://www.youtube.com/embed/eVX7c_eviB8?rel=0

This wonder of neurotechnology will cost you only $299 USD a basic SDK starts at $500 USD for those who wish to develop custom software for it. Emotiv already feature demo applications for their neural
interface but the possibilities are limitless. No doubt this and similar interfaces have their limitations (access
surface, gel application, background noise and interference) but are bound to advance further as wide-spread
commercialisation takes place, allowing for market-funded research to take place.

It is not hard to imagine how this technology may apply in search, usability, emotional response testing,
advertising and marketing. It’s only a matter of time before somebody writes an app to search Google - by
simply thinking about it.

Let’s indulge in speculation on the (distant?) future of search. Leave your questions, links and ideas in the
comments field.

References:
_singularity on 17 October 2011.

Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, Jack L. Gallant - Current Biology - 11 October 2011
(Vol. 21, Issue 19, pp. 1641-1646)

on 17 October 2011


3.6.2 Link Building Software (2011-10-28 02:29) - admin

This study focuses on analysis of link building workflow and delivers a practical framework for process
optimisation. Research activities which lead to this study were conducted in the period between January
2009 and June 2011 by Dan Petrovic and a team of no less than twenty link builders at Dejan SEO.

The answer to most of our challenges is presented in the conclusion of this article. Read on however for a
journey into anatomically dissected link building process.

How SEO Companies Approach Link Building

One of the most common mistakes SEO companies make when engaging in link building activities is measuring
success by link quantities. It’s important to remember that the web is an infinite source of links and the
pace at which new documents are created is greater than that at which you can create links - even if you use
software and automation. The real commodity in this situation is your time. Allocating time appropriately is
the key to optimal performance of your team.

Assessment of Research Styles
There are many effective methods of determining which links are valuable and which ones are not. There are a few practical methods commonly used by SEO professionals involved in link building:

- Software Link Analysis
- Manual Link Research
- Real-time Link Opportunities
- Existing Link Profile

Software Link Analysis

The purpose of software link analysis is to save time by automating activities not practical for research by individuals. Let’s explore a few facets of software based link analysis:

Topical Link Analysis

This step is not as complex as it may sound and is based around initial topic parameters which initiate software’s path. There are three parts to this process:

1. Initial Parameters (e.g. Keyword, Specific URL)
2. Crawling & Recording Quality Parameters
3. Sorting of Data

Given initial parameters, crawlers starts the harvest process following keywords and links. Keep in mind that after third or fourth level the topical dilution may be too great to justify continuing research - that would be a good stage to stop the crawlers and initiate qualitative analysis.

Competitive Link Analysis

Basic competitive link analysis can be performed on two levels:

1. Single Competitor
2. Multiple Competitors
In the first case the main objective is effective sorting of qualified links. In the second case there are additional considerations such as:

1. Data Merge
2. Duplicate Removal
3. Link Overlap Analysis

Harvested links can be merged into a unique list free from duplicates and added to the link overlap database. Overlapping links indicate easy targets where likelihood of successful link placement is high. Using this method in combination with topical analysis allows discovery of relevant and likely link opportunities.

Qualitative Link Analysis
When deciding which links to acquire first it is important to understand the links’ potential. This is very difficult and time-consuming if done manually and is another example of activity which can be effectively replaced by software. List of qualitative parameters may contain the following:

- Domain Strength (e.g. Domain Authority, Inbound links from gov/edu/mil)
- Page Strength (e.g. PageRank, MozRank or ACRank)
- Incoming Backlinks (Quantity & Quality)
- Outgoing Backlinks (Quantity & Quality)

Opportunistic Assessment
Before targeting a specific link opportunity it is wise to understand the likelihood of that link being placed on the selected page or website. During qualitative analysis we seek pages which do not link to many other external domains, but in some cases existing outbound links may be an opportunity signal. A page which contains no existing links is less likely to accept a new link than the one which already contains some.

Opportunity assessment can take place on two levels:

- Domain Level
Additional filtering option may prevent false (or very unlikely) opportunities:

- Nofollow (link, meta, robots)
- Exclusion of commercial results (e.g. only links to gov/edu/mil domains)

Based on this logic we have developed a tool which highlights link opportunity pages by scanning a single domain of our choice and mapping the pages with existing outgoing links while filtering with above mentioned parameters.

Manual Link Research

Manual research typically brings the most original opportunities, but often in insufficient quantities. This is why many SEO companies avoid this method and give priority to more automated solutions which do not deliver sustainable results and may even be in breach of search engine guidelines. In order to minimise the time consumed by manual link research it is important to analyse and understand the link builder behaviour and deliver a set of productivity tools which will fast forward the process and enable streamlined workflow.

Linear Path

Linear research is based on a single browsing path and represents the most streamlined of all research methods. It typically manifests in three distinct styles:

- Interrupted
- Cumulative
- Hybrid

Interrupted linear link building style requires link builder to act on each opportunity as they encounter it. Cumulative is based on collecting opportunities on a linear browsing path and acting on them after research path has been exhausted. Hybrid is the combination of the two where link builder decides which opportunities are to be acted on immediately and which ones are to be accumulated for delayed action. The main challenge in this research style is simply quick and efficient storage of data and ability to track the research path.

Diverging Path
Based on linear link research, diverging style involves branching out into multiple research threads. This style yields broader and richer data but is more challenging due to limitation of our ability to store and retain information. Researcher will typically end up with multiple browser windows and tabs open and attend to each one in a linear fashion.

Regressive Path

Both path styles require easy access to mapped activities and ability to track back to the original research node. This is particularly useful if the original node diverged from the very start.

Stage Analysis

At the end of each research stage, link builder may wish to review the progress on a holistic level where an overview screen might prove useful. Stage overview represents a collection of URLs sorted by selected criteria (quality metrics, topic, and link metrics).

Real-time Opportunities

Link opportunities occur in real-time and in random intervals with absence of research path and methodology. Examples are:

- It is acceptable to evaluate each new instance manually but integration into existing link management system with alerts and notifications can indeed be helpful.

Existing Link Profile

Analysis of the existing backlink profile is another effective way of discovering potent link opportunities through:

Old and well-established domains tend to have thousands of existing links which could be further optimised by correction of the linking environment or the link itself. Presence of links on other domains may point out at a better placement method on the same domain and spark ideas for new link locations.
Extraction of existing backlink profile is possible using Google Webmaster Tools data and for refined dataset optional embedding of Google Analytics or other statistics packages into the mix. Observed elements are:

- Anchor text
- Image links
  - Use of ALT text
- Landing page potential
  - Number of outgoing external links on the page
- Noindex / Nofollow

This can be applied on a top level for a merged dataset of multiple clients. This is an excellent way to determining common link hubs which can be made into basic link opportunities for all new projects.

**Human Interaction & Interface**

While software-based link analysis produces useful lists we found that manual link research does not flow well enough to enable smooth research. Traditionally link builders use spreadsheets to enter link opportunities with varying degrees of sophistication and sorting ability. The shortcomings of this method are obvious in a collaborative environment where more than one link builder needs to access the stored information.

The following diagram illustrates what happens to each link opportunity and how software automation helps with data entry.

Software can also take care of many miscellaneous link properties which have to be entered in previous versions of our link building framework. Here is how a link opportunity may transform in time:
Streamlined workflow has its consequences and much is being sacrificed in terms of data accuracy. This is why it’s important to enable optional second step and post-entry data population with easy skipping option.

Main link database interface is complex and contains numerous additional functions so the real challenge is to enable flawless user experience in order to minimise administrative time and increase productivity. Before going into our interface solution let’s break down the properties of a link to get an appreciation for the complexity of link-based inventory system:

- Full Size PDF: [1]Link Anatomy Visualisation by Dejan SEO
Illustrated below is the wireframe interface designed to incorporate the most essential options for daily workflow:

Two key areas are the search and refining (sorting & filtering) which enable link builders to quickly find the most relevant opportunities and prioritise their contacting campaign. This feature becomes increasingly useful with time, especially in larger teams.

Dashboard
Dashboard serves as a quick overview of the latest link building activities such as link status updates and alerts. Dashboard can also be used for team-wide notifications and ranking updates. We found that link builders prefer to see regular ranking updates as it provides meaningful connection of their work to ongoing success of the campaign.

My Projects
This section provides in-depth information on link building activities for each specific project, allowing team members to monitor progress in rankings.

Link Digger
Semi-intelligent link suggestion area based on tagging, topical relevance and link likelihood.

My Links
Area where link builder can review their progress and links scored in the past.

Favourites
During link search within the system link builders can highlight links they like the best and quickly find them in their personal favourites list. Items which get highlighted by multiple users progress in the top chart of all-time favourites.

Import Links
One thing we’ve learnt during our study is that no matter how good the system is, people will always use offline records either for speed, exclusivity or backup. Having import/export functionality is a must.
Link Exclusivity

An interesting phenomenon gradually develops in large link building teams where each member of the team creates a personal list of contacts and links they do not wish to share with others. This is partly as those links are perceived as a personal secret weapon and partly to protect a well-established relationship from external intrusions (even if they are the members of the same team). Within our team this has been solved through an "exclusivity tag" which mark a link as proprietary (specific either to project, contact person or a member of our team).

My Account

Apart from basic account information this area is also designed to host success trophies and awards which everyone else can see.

Tools

Central collection of internal and external tools which can be used for advanced research tactics and manual link research.

Team

Browse team members and their links, see who is the best performer.

My Notes

Ambitious attempt at replacing offline notes and records. Does not get used at all.

The Final Solution: Introduction of the Browser Toolbar

After numerous attempts at creating and improving interfaces it was decided that a browser toolbar will provide the best possible user experience. How toolbar works is similar to adding items to an online shopping cart or bookmarking pages in a browser.

Click the "add" button on every page you think you can, or would like to get a link on. Link builders can now quickly flag and collect link opportunities without being taken away from their research path. This is particularly useful if somebody already managed to score a link on a particular domain and others can see this in the toolbar alert and read the instructions on how it was done and who to contact. The toolbar interfaces with the existing link database and is able to write and retrieve link status in real-time.

What’s Next?

Our link platform is currently in its alpha stage and rigorously tested for best possible user experience. If you are a link builder and have a burning desire to contribute or to test-drive [3]please get in touch and let
us know. We’re expected to go live with a public beta in 2012.

Credits go out to Matt Morphett, one of Australia’s most talented user experience experts for his continuing contribution to development of our framework. We would also like to thank people at SEOMoz for their excellent metric data.

How to Apply for Alpha Tester Access?

Please fill out the following form:

[4]https://docs.google.com/spreadsheet/viewform?formkey=dHRlcjcw dRqLTh2ZXNLMHBBVkdPa0E6MQ

3.7 November

3.7.1 Applying Random Surfer Model to Peer-to-Peer Network Distribution  
(2011-11-02 20:22) - admin


Digital information preservation is a hot topic and a fertile ground for many bubbling solutions and models in both practice and theory. One of the emerging issues revolves around the fact that there is more information being produced today than we’re able to store and analyse, not to mention attempts at prioritisation and archiving.

[2]Kate Cumming (Future Proof - State Records, NSW) writes about this issue in her article titled: [3]’Here comes everybody’: What does information ubiquity mean for the archives?

Two issues in particular caught my attention:

"The gap between what we create and what will actually have the capacity to keep is also growing exponentially. ‘In fact, the production of digital information has already outstripped global server capacity by an estimated factor of four or five’"

"We are also yet to fully define and understand the true costs associated with archival storage in this digital world."

It seems clear that if we’re to reduce administrative costs of information management we must minimise (manual) human intervention and secure continuity through robust and timeless protocols which are time, place, machine and platform agnostic. What about the problem of rapid information production? How can we process (capture, classify, verify, validate and order) the incredible amount of new information being
generated every second?

Assessment of Existing Technologies

Origins of the basic logic for PageRank can be traced back to 1940s [3] and various scholars and scientists thought about the problem in various periods of the 20th century. The proposed algorithm observes a collection of documents and analyses how they link to each other, counting each link towards a page as a vote. Google uses this technology even today, though on a more sophisticated level and in combination with other reputation signals. One can argue that documents which earn more trustworthy links and mentions tend to be more valuable than the ones nobody talks about. Although not bullet-proof, this method is capable of moderation of vast quantities of documents without any human intervention.

Could a similar voting system be applied in digital preservation frameworks to flag document in a digital collection as more valuable on the basis of its references?

Is Existing Technology Applicable?

After discussing the idea with Adrian Cunningham, the director of Digital Archives Program (Queensland State Archives) it was brought to my attention that the application of this algorithm may be more useful on an aggregation- rather than document-level.

"Archivists, for the most part, manage records in aggregations and not at the level of the individual document - that is partly because there are too many documents and too few archivists, but it is also because archival philosophy assesses the evidential value of records in their context and their context is largely derived from the aggregations of records of which they are a part and the inter-relationships between records within those aggregations."

Another very good point is that archivists assess the evidential value of records from their context rather than their informational content. For example:

- Why was the record created?
- Who created it?
- During what activity?
- In which sequence of events or transactions?
Content analysis algorithm may not prove so useful unless they can extend to cover context and although some context can be discerned through content, this provides no reliable method on a broad level. Algorithmic treatment may work on a slice of the documents which tend to already have good recordkeeping meta data associated with them as that would provide good documentation of context. To make the situation even more difficult we can still add the challenge of standardisation of meta data to provide comprehensible and reusable format.

Creative Storage Solutions

Let us for a second imagine that by some magic the sorting and prioritising part no longer represents a challenge. What other problem do we face today? Storage of course.

Many talk about the cloud as an ideal solution, although in my view, it doesn’t really bring anything revolutionary into distributed computing model - other than a catchy name. Many successful distributed storage platforms have evolved in the last decade and most of them were used for illegal purposes. If we are to learn one thing from movie and music piracy is that P2P (Peer-to-Peer) networks do provide robust and resilient storage solution. Many of these protocols (such as torrent) exhibit helpful features such as integrity checking, redundancy, fragmented storage, basic resource prioritisation, availability indexes, trackers and various other meta data.

At first the idea of algorithmic treatment of archived collections in P2P environment seems a bit farfetched, but there are some interesting developments in that area already. One of them is the [6]The LOCKSS Peer-to-Peer Digital Preservation System [6] which stands for (Lots of Copies Keep Stuff Safe). LOCKSS is an international community initiative with the objective of securing affordable preservation of digital content for libraries. They have been running for over a decade now and thanks to their open-source approach, developers are able to aid the evolution of supported platforms and applications.

The question remains, could a sophisticated algorithm with enough contextual data self-organise large-scale document archives in a distributed storage network without any human intervention?

References:


3.8 December

3.8.1 Introducing Phrase Potential Calculator (2011-12-05 15:45) - admin

[1] Phrase potential calculator is a tool which analyses CTR, rankings and traffic volume data and helps discover phrases of high value. The calculator uses data from Google Webmaster Tools and/or Google Keyword Tool.

What you see here is keyword output from a report generated through traffic data sourced via Google Webmaster Tools CSV file and manually added goal conversion value and conversion rate.


Downloads:

Download this article in PDF format: [7]Applying-Random-Surfer-Model-to-Peer-to-Peer-Networks


©2013 dejanseo.com.au/blog
We have a rule at Dejan SEO; when something takes more than one hour to do and you do it often, we write a script for it.

Notice the first term from the list. "Electronic music" has local search volume of 50,000, and with #4 rank in Google it receives 1600 clicks. This is already a fair bit of information and we know the CTR value from SERPs so we’re able to model and create scenarios (e.g. Clicks in #1, #2, #3 or one position up). We are able to calculate CTR averages for this domain through its own values or average out if the data is insufficient to work with.

At the bottom of the report our calculator gives a summary of total traffic and for each new ranking scenario as well as totals in projected revenue. Graphs add visual candy to the report making it suitable for meetings, management briefs and decision-making visualisation.
There are two methods of calculation: Phrase Specific and Average Based. The difference between the two is that one uses CTR averages throughout the site (or uses across the board averages) while phrases specific calculation observes each individual phrase and projects based on performance track of that phrase. This is particularly useful when you have a lot of branded terms which tend to outperform generic phrases and can skew the CTR in favour of a first result (or first three in many cases).
"Average Based" calculation allows you to modify CTR values manually while "Phrase Specific" calculation retrieves CTRs for each phrase individually:

To add goal conversion value and rates click on the traffic value button in the financial potential calculation section:

In a nutshell this tool helps you decide which search terms are worth targeting and how much more money you could make if certain phrases were to perform better in search results. The potential calculator becomes very helpful once you start dealing with large quantities of data. The tool is still in beta and we’re collecting feedback from users at the moment.


3.8.2 Synonym Anomaly in SERPS (2011-12-12 09:50) - admin

This is a transcript from a Google+ discussion on a recent anomaly we spotted in Google search results. Two separate but obviously linked terms ("temporal" and "dreaming") are bolded when adding the word "plot" afterwards. AJ Kohn (blindfiveyearold.com) and Bill Slawski (seobythesea.com) provide some insightful hints as to why this might be the case. Steven Baker from Google joins in acknowledgment of the case which has
been brought to attention of relevant groups within Google’s query understanding team.

Question for +AJ Kohn:
Why does Google highlight both "temporal" and "dreaming" in this search query?


AJ Kohn - Oh, this is an interesting one. It’s not a straight up synonym since dreaming and temporal are not returned as normal Google synonyms. The 'temporal' substitution doesn’t happen when you simply search for 'the dreaming void'. However, when you look at the related searches for this term you’ll see 'temporal void' as the first related search. So, Google understands there’s a relationship here between these terms based on query patterns. The 'plot' modifier seems to trigger an even higher degree of relatedness (though it isn’t shown in related searches). So Google believe the two are similar enough to warrant returning (and highlighting) both terms. You can sort of test this by using the new verbatim search which does then remove the 'temporal' results from that query. So I’d say this is a type of modified synonym based on a high probability of relatedness. Mind if I blog about this? 00:01 +1

Wissam Dandan - +AJ Kohn do you think this is relevant to dan’s search "Related query results refinements: Sometimes we fetch results for queries that are similar to the actual search you type. This change makes it less likely that these results will rank highly if the original query had a rare word that was dropped in the alternate query. For example, if you are searching for [rare red widgets], you might not be as interested in a page that only mentions “red widgets.”” [2]http://insidesearch.blogspot.com/2011/12/search-quality-highlights-new-monthly.html 03:07

AJ Kohn - Well, yes and no +Wissam Dandan. The related search query refinement is what is going on in this instance. But because the query didn’t have a dropped word, it still returns all the results. What’s interesting is that it is bolding both the original term and the related term. I’m not sure I was aware that they would do that in these related query results. Frankly I like the term 'query synonyms' myself, I think it’s a bit more easier to grok. Good find BTW. 03:21 (edited)

Dan Petrovic - Blog away +AJ Kohn that’s why I sent this 'anomaly' to you - you’re a man who appreciates Gattaca and no doubt has a good insight into science fiction. One person I forgot to ping in on this is +Bill Slawski - let’s see if he can clarify the strong connection of these terms through his own knowledge of the subject. 07:34

Bill Slawski - The series sounds interesting. Would you recommend it? I accidentally killed my first shot at an answer to this post, but I’m going to try again. Will break this up into a couple of posts to help prevent that. There are a number of Google patents that address semantically related terms for query expansion, though most of them do so with the context of queries that contain more than one word. For instance, while we may often consider the words "auto" and "car" to be synonyms, that’s not the case when you set an alarm on "auto." Even within longer phrases, words that we might consider to be synonyms might not be. So,
"automobile" and "car" are synonyms when we search for a [ford car], but not when we search for a [railroad car]. 08:20 +1

Bill Slawski - Google patents that I would suggest looking at include this early one, from 2003:


The next batch all seem related in aim, but provide some different methods to reach that aim:


There are a couple inventors names that stand out somewhat on these patents. +John Lamping is listed as a co-inventor on the first three, and +Steven Baker is listed on the last 5, but I don’t expect them to show up and give us some hints here (though that would be really nice.)

For looking at the query terms, the query pattern [the _ _ _ _ _ _ void], and the fact that many terms probably co-occur on pages about them on the Web because of a shared universe, related storylines, the same author, and possibly even shared characters, I’m not completely surprised that they appear as semantically connected query expansions. It would be worth looking to see if this happens with the titles of other books from authors who used the same structure in titles to his or her works. 08:46 (edited) +4

AJ Kolm - +Bill Slawski, mind me snagging a quote or two from this thread for a blog post? 08:32

Bill Slawski - No problem, AJ

AJ Kohn - Thanks +Bill Slawski.

I find it interesting that they bold the query synonym terms. As an aside, I don’t see this behavior with Kim Stanley Robinson’s Mars series (Red Mars, Green Mars, Blue Mars). 08:46

Bill Slawski - Thought of looking at Robinson’s Mars series title, but decided that it’s possible that the term "red mars" appears in many other contexts as well. Not too much luck with some of the Dune books, either. 08:54 (edited)

AJ Kohn - Nothing for Donaldson’s Gap series either (i.e. - The Gap into Conflict/Vision/Power/Madness/Ruin). 09:02

Steven Baker (Google) - Thanks for bringing this to my attention. I can’t comment on this in great detail, but I have brought these examples to the attention of the relevant groups within our query understanding team. BTW, here’s a blog post where we talk more about our synonyms techniques in Google search: [10]http://googleblog.blogspot.com/2010/01/helping-computers-understand-language.html

AJ Kohn - Thanks for joining the conversation and for the link +Steven Baker. Any details you could provide would be greatly appreciated, though not expected for obvious reasons. Perhaps an innocuous one you can confirm? Is it normal for the ‘query synonym’ (for lack of a better term) to be returned in bold? [edit] It looks like that’s actually answered in that blog post!

Historically, we have bolded synonyms such as stemming variants — like the word "picture" for a search with the word "pictures." Now, we’ve extended this to words that our algorithms very confidently think mean the same thing, even if they are spelled nothing like the original term.

Thanks again for your time and engagement.

10:47 (edited)

Dan Petrovic - Thanks +Steven Baker - the query is certainly strongly linked but not necessarily useful to my search at the time. I personally consider it a borderline bug.

Steven Baker - +AJ Kohn, in the blog post I linked to, we say, "We also recently made a change to how our synonyms are displayed. In our search result snippets, we bold the terms of your search..... Now, we’ve extended this to words that our algorithms very confidently think mean the same thing, even if they are spelled nothing like the original term." +Dan Petrovic, that’s good feedback for the team. We don’t fix specific problems but we use such examples as feedback to improve our signals and algorithms.

AJ Kohn - Yup, just found that quote +Steven Baker. Thankyou. It’s an interesting line between being highly related and a true synonym. Oddly, both could help improve search quality to some degree, though I’d agree with +Dan Petrovic that in this case the added ‘query synonyms’ don’t improve quality.
Bill Slawski + Steven Baker

Thank you for joining the conversation. I remember your blog post on language, and have provided links to it more than a couple of times in the past when people asked me about synonyms, as well as this one involving language models: Making search better in Catalonia, Estonia, and everywhere else [11] http://googlepublicpolicy.blogspot.com/2008/03/making-search-better-in-catalonia.html

Those types of posts are really appreciated. I’m really happy to see that you’re actively engaged in receiving feedback on synonyms as you noted in the first blog post. I actually like having these related results show up for these particular queries (the dreaming/temporal void, and Soldier of Arete/the Mist), and have a sense of why they are, but I’m not sure if they should show up at mixed in with the other results, but maybe instead as related results. I’m wondering how often results that are so semantically related, yet refer to different entities end up appearing within search results. Again, thanks. I listed a number of Google patents related to synonyms and query expansion above, and after rereading the Catalonia blog post, I should probably include this one as well:

Machine Translation for Query Expansion

Contributor Credits & Links


Original Thread: [15] https://plus.google.com/111588754935244257268/posts/iLnVLmNyBGK

15. https://plus.google.com/111588754935244257268/posts/iLnVLmNyBGK
After many years of demand by the SEO and webmaster community, Google has finally given us an API option to download search query data from Google Webmaster Tools.

For years now, the drill has been the same. Log into Google Webmaster Tools, navigate to search queries, set the parameters, date range and filters and export data. We were grateful to have an export function in the first place but the way to download data was very restricted and it still is for many things including link data. Now we have an option and it seems they are busy with enabling more. [1]Jonathan Simon, Webmaster Trends Analyst reports on this in detail in Google Webmaster Central [1].


To make use of this facility go to [3]Google Data APIs Python Client Library download and install the library. Add the script [4]downloader.py to a new folder and copy the following script [5]example-create-spreadsheet.py to it as well. (Customise it to replace the dummy values). Run the example-create-spreadsheet.py script in your Terminal:

```
python example-create-spreadsheet.py
```

You can see your data in Google Docs:

![Google Docs screenshot](image)

Use [6]example-simple-download.py to download the spreadsheet (CSV) without storing it in Google Docs. What’s great about this script is that it can be configured to regularly download and store fresh data ready for manipulation, reports and visualisation.

Watch out for security as this script contains Google account username and password stored. The script uses https as default as the first layer of protection of your data.

References:

[1] Simon, J. - Google Webmaster Central Blog:

1. [https://profiles.google.com/jonathansimonweb/](https://profiles.google.com/jonathansimonweb/)
Chapter 4

2012

4.1 January

4.1.1 Define: "Product" (2012-01-22 01:00) - admin

Even in the pre-Panda SEO world faceted navigation was a double-edged sword. Today though, if you let this monster rule your website bad things can happen - on a mass scale. How can webmasters understand how their website architecture truly looks like when they struggle to visualise it?

The above image is a representative slice of a very large product-based website. The visualisation was created using PowerMapper’s Sitemap Generator which was the best application for the job at the time. The illustration above is incomplete and a point at which the software stalled and the crawl was interrupted manually. Why? Because the website architecture lead to infinite combinations due to various facets, browsable filters and tags. Here are a few common factors to mention:

- Canonicalisation (solved with rel canonical)
- Pagination (solved with rel prev and next)
In another attempt we pick a smaller section of the website and let our software run longer to see if the "infinity" theory really holds and it seems to have done so. After hours of scanning we finally give up, producing the following visualisation:

So what can SEO professionals do? Well, there are industrial-strength computation solutions available for big bucks (IBM, Microsoft, Oracle, Fuji Xerox, Siemens...etc) and there are a few firms with clever internal
solutions. At SMX Elite in Sydney Dennis Goedegebuure showcased examples of robust-looking custom-built software at eBay (though not strictly related to architecture mapping). For now I would say all hope is lost for visualising and structuring large-scale hyperlink data within a single domain (unless you are Google). Instead what we can focus on are the best practices in terms of keeping website’s structure lean and focused and content as rich and unique as possible.

Sometimes it’s about drawing the line and deciding what is a product and what isn’t:

Look at this fine example of 90’s design Reebok shoes. What defines this product (other than pure awesomeness)?
Primary Properties

Brand: Reebok
Type: Insta Pump
Model: Fury

Without the above you could be looking at any shoe type and not find what you are looking for.

Secondary Properties

Gender: Male
Size: 10
Price: $135

Tertiary Properties

Colour Scheme: "Green" by Tyler Gruwell
Hexalite Air
Graphite
Pump
Reflectors

Question: Is this a different product to the one above?
The only difference is the colour scheme "Versuch 2" by Möhring Denis. Well, Reebok seems to think so and they let users submit their own designs, in great numbers. At the time I write this article there are 469 different colour schemes for this one shoe type on Reebok’s website, all user-generated and promoted via social media. The concept is fantastic, but arguably a smaller website could probably not get away with this stunt, not after Panda and not without social signal validation.

If you take a look at the source of the page you will see that all these shoes are set to both INDEX and FOLLOW. Reebok did one smart thing though, there is no faceted navigation in the mix. If they were to
implement that the number of indexable pages would skyrocket from 500 to 5,000,000 and it would be only a matter of time (and PageRank) before those pages are in Google’s index.

That would surely trip the Panda filter.

Going back to our original product definition dilemma, let’s say it’s safe that a product is defined by its core parameters and surrounded by secondary and tertiary characteristics. Different products (and markets) may require some customisation but in our case it’s quite clear that certain properties should never form a new product.

An example of this would be shoe size. It would be completely silly to create two pages for the above shoe model, one with size 10 and the other with size 12. The available sizes are present on the page and will index and be found for in search.

The Problem

Oh, but it used to work so well. One would create product pages for every possible variation of the product including all non-essential properties such as shoe size. This lead to 10,000,000 page websites and bloat in
Google’s index and resulted in a birth of little Panda filter whose job is now to trim the size of the web.

What’s best for the user?

Oh here we go, we’ve heard this phrase so many times from everyone at Google. The truth is that they don’t really know. In last-year’s analysis of eCommerce search queries I found that Google seems to be magnetically attracted to the bottom line pages (e.g. exact product instead of a category). They always strive to give the user ”the end result” and minimise the search. Oh, how many times they get it wrong, it’s incredible. This happened so much that it started reminding me of [3]Clippy. Sometimes it’s useful to go over categories, especially in the early research stages - being shoved into a random end-product is not helpful.

Perhaps the solution is in expanding the old definition of the product:

Into something a bit more elaborate:

Have you seen any examples of good website architecture and well-defined products? I'd love to see it.

Please comment below.


4.1.2 Post-Panda News Results & Search Quality Assessment (2012-01-26 12:53) - admin

After [1]Panda consumed much of the [2]bamboo on the web, [3]MuffinTop algorithm update was supposed to mop up the rest. There are still a few 'evil unicorns' frolicking around in the flowery meadows of Google's search results. In this article I describe two of them - one from standard search results, the other from the news.

Case I:
Believe it or not, it all started with people laughing at my old Samsung CRT TV. Under peer pressure I go and buy the biggest and best smart TV I could find. This left me with an old box that I could not get rid of so I decided to sell it on eBay. This is where my search starts. In hope to find the manufacturer specifications I type in the model number in Google: Samsung TV CS29K30MG and click on the third result. The website I get is filled with ads, has no content and asks me to join to get my information. Naturally I'm furious. 

I hop on Google+ and ask the question: "I want to know why is this piece of rubbish website ranking in Google after Panda and Page Layout Algorithm. Anyone? Enlighten me!"

A good debate starts (thanks to Sach Mayer and Lyndon NA) on why this website slipped through the recent algorithm upgrades. And the initial thought is that there were simply not enough search results and Google had to show something, no matter how useless.

Later on John Mueller starts up a Google Wembasters Hangout and Matt Cutts joins in from India, so I asked them to clarify and explain what happened.

Matt calls this type of search result "an evil unicorn" and confirms our suspicion that there simply wasn’t enough to show and Google had to display something.

One good point Matt Cutts makes in the hangout is that Google should perhaps display a message that says something like this:

"We have found an unusually small number of results for your search query. In order to improve the quality of the information you find try refining your search using different combination of keywords."

I think that’s a great idea, as not all users are comfortable at power-tweaking their results to refine the quality of search.

The debate continues on Sach Mayer’s Google+ post in which Pelagic, Lyndon NA and John Mueller analyse another "in between" case where quality content with plenty of links exhibits characteristics of ad-loaded low quality sites but gets away with it even in a moderate search result number type query (Rabbie Burns). One element this page does have is the exact match domain, relevant to the search query which would have additionally boosted the result.

SEO Takeaways

1) Panda & MuffinTop page layout algorithm don’t really remove or filter out any search results. They are
still there, buried somewhere at the bottom of the sorted list of results.

2) You can still get away with ads if you are authoritative enough or if you have a domain that is seen as a good match to the search query or a brand search.

3) Very specific terms will with minuscule search volumes and low number of results have a chance at ranking even on thin content pages full of ads. Keep in mind that although these may rank they can still affect the rankings of your better quality content so this is overall not a good strategy.

Amit Singhal writes:

"One other specific piece of guidance we’ve offered is that low-quality content on some parts of a website can impact the whole site’s rankings, and thus removing low quality pages, merging or improving the content of individual shallow pages into more useful pages, or moving low quality pages to a different domain could eventually help the rankings of your higher-quality content."


Did you know?

Until recently you could not find Google+ by searching for it in Google. This was due to the fact that symbols and punctuation would normally be ignored and omitted. In the above mentioned Google Hangout Matt Cutts states that the update only coincides with the release of Google+ and was not introduced to help their social media platform be found in their results.

Case II:

This morning my phone alerted me to a ranking change for "SEO" so I went to [17]investigate and found an unusually poor news result merged with the standard search:
Search snippet was bad enough but when I clicked on the link it looked like this was only a verification test to see if the Toronto Star would render well into Google News Search. OK, fair enough - bloopers happen.

Have a look at the layout of this page though:

OK, the content is missing - but what else?

How about I colour code this layout for better visualisation:
So first you get the ads, then you see internal partnership promotions followed by logo, search box and fluff useless to user looking for content, this could have in my opinion been reduced to a shorter block. After that a standard layout element (navigation) is visible which is common enough - nothing wrong with that. But what bothers me here is that instead of actual content we see more internal promotion (other articles) followed by some more ads on the right.

Content? Nowhere to be found.

Check the [cached version of this page](https://www.google.com/cached?dws=1218133216141372000&dqt=2&dq=imagine%20you%27re%20looking%20for%20content%20but%20have%20to%20fight%20past%20all%20the%20ads%20on%20this%20page) (update: the page is no longer cached) and you will see it’s in (It is a snapshot of the page as it appeared on 25 Jan 2012 21:02:29 GMT.)

Something is definitely wrong with this picture, considering many great pages have been accidentally filtered out with the recent algo changes while high authority sites get away with thinner than thin content (read: zero) and plenty of ads.

Google has done a great job at removing the rubbish from its bloated index but the next step should be to fine tune towards a universally fair algorithm.

[styledbox type="general"]Update: Update: John Mueller from Google has informed Google News Team about this result.[/styledbox]

We would like to hear your comments and opinions on the Google+ post here: [19]https://plus.google.com/114074532743058808065/posts/EjyCZd1Xj5c
4.2 February

4.2.1 Missing Title Tag Substitute (2012-02-01 00:19) - admin

Search for [1]intitle:"Untitled Document" in Google and you will find plenty of websites with missing title tags. Worry not though as Google looks for hints of the page’s main theme elsewhere in the document and inserts it into search results instead of the title. So which elements can be used as a substitute? I investigate further.

Update: We’ve renamed the title of this very article to ”Untitled Document” and the following is Google’s re-write based on our H1 and brand name:

Summary of Findings

Which elements will Google pick to replace a missing or inadequate title tag?

1. Domain Name
2. Page URL (note: gets confused with multiple levels)
3. Domain + URL combination
4. H tags
5. Plain Text (tested when at the beginning of a document)

6. Elements of a parent page (in case of iFrames)

7. Truncation for boilerplate against variable elements.

Untested but highly likely: Anchor text in inbound links, DMOZ, local citations.

Our Research in Detail

Case 1: http://www.sarahaking.com/

Rendered Title: Sarah A. King

Inference: Likely <font color="#006699">SARAH A. KING</font>

Note: Domain name could have been used but it doesn’t contain punctuation visible in the rendered title.

Case 2: http://www.wesleyburt.com/drawingsamples.html

Rendered Title: Drawing Samples - Wesley Burt

Inference: {URL_filename} - {Domain Name}

Note: Potentially also home page TITLE.

Case 3: http://www.pinholephotography.org/Solargraph %20instructions.htm

Rendered Title: Solargraphs - Pinhole photography

Inference: {URL_filename} - {Domain Name}

Note: Identical inference to the one above.

Case 4: http://stepheneastwood.com/tutorials/lensdistortion/strippage.htm

Rendered Title: Stephen Eastwood

Inference: {Domain Name}

Note: /1/2/3/ levels of directories may be the cause for the drop of this element in the title substitute.

Case 5: http://www.howardtangye.com/HT-90.html
4.2. February

Rendered Title: Untitled Document - Howard Tangye

Inference: Untitled Document + {Domain Name }

Note: How interesting! Google decides to keep the Untitled Document part and use only domain but in addition rather than completely replacing it? Do they think the piece is actually called "Untitled"? Perhaps not but a semantically challenging HT-90 may have something to do with them giving up on figuring out what it may be.

Case 6: http://www.benkler.org/CoasesPenguin.html

Rendered Title: Coase's Penguin, or Linux and the Nature of - Yochai Benkler

Inference: H3 - H4

Case 7: http://www.scientificamerican.com/media/8-Wonders/01-Intro.html

Rendered Title: 8 Wonders of the - Scientific American

Inference: Parent Page Fragment (iframe) - Domain

Note: This is a strange case in which Google gets it wrong and truncates the title. What’s interesting is that Google follows the [2]parent page in which this URL is embedded as an iFrame and gets its information from there. My guess on truncation is because this goes into several sections each slightly changing (e.g Saturn Rings).

Case 8: http://www.tonyhawk.com/thth/

Rendered Title: ( #THTH) Rules! - Tony Hawk

Inference: H2 Fragment - Domain

Note: This is a case where truncation is useful and Google removes general site title and leaves only the bit of substance (e.g. Tony Hawk’s Twitter Hunt ( #THTH) Rules!)

Found any more interesting cases?

[3]Let us know on Google+
In a recent [1] SEO experiment we tested several cases of title tag replacement in Google and discussed elements which are used for replacement and when it may happen. Today we found another great example with some additional insights as to how Google may perceive information on the internet from the semantic point of view.

Let’s observe the good old branded snippet:
Nothing interesting there as our title tag is as it displays in the search result. From time to time Google may pick brand name alone in case where title is saturated with promotional terms and keywords.

In our next case we see something quite interesting when searching for [2]Sennza in Google:

Same situation with [3]mobile results. I thought this was strange and went to investigate:
Luckily it was dead obvious what caused this, but it did not make it any less interesting. So if you say "we" Google may associate the next word with the group, organisation or brand which is interesting and makes a lot of sense. There is plenty of cases where that may not be the case, but as we can see Google got it right this time. What is odd about this case is that they left brackets around the brand name and even inserted it in the search snippet instead of their title tag (which is optimised for SEO):


In this case Google though, oh well, you searched for Sennza so I think it’s better to show Sennza in the title than the above. Why they used brackets around the brand is a mystery to me and looking forward to hear [4]opinion of our readers on Google+. If you like this sort of thing go and read more about our recent [5]title tag experiment which lead us to following conclusions about elements used in title replacement:

1. Domain Name
2. Page URL (note: gets confused with multiple levels)
3. Domain + URL combination
4. H tags
5. Plain Text (tested when at the beginning of a document)
6. Elements of a parent page (in case of iFrames)
7. Truncation for boilerplate against variable elements.

Update: Google Webmaster Tools picked up this page title and classified it as "not informative" and presented it as one of the site issues as visible in the screenshot below.
### Non-informative title tags

Your title provides users and search engines with useful information about your site. Text contained in title tags can appear in search results pages, and relevant, descriptive text is more likely to be clicked on. We recommend reviewing the list and updating the title tags wherever possible.

*Back to all issues*

#### Pages with non-informative title tags

<table>
<thead>
<tr>
<th>Pages with non-informative title tags</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Untitled Document</td>
<td>/missing-title-tag-substitute/</td>
</tr>
</tbody>
</table>

### 4.2.3 Link Graph Visualisation (2012-02-17 19:15) - admin

Dejan SEO team is working on link graph visualisation methodology in order to develop link and social network analysis framework aimed to solve risk assessment challenges, influence calculation and social graph analysis.
The big bright dot is Dejan SEO, scaling down in colour intensity with each next link connection. Observe the small groups of interlinked websites - this is one way search engines can determine how one website related to another and also use for link spam detection.

The first step in link graph visualisation is of course data retrieval using custom crawling with strict parameter limiting. The output file produces a large but simple CSV with link source, link target, linking domain, linked domain and anchor text. No other metrics are used at this point in time, however we are considering applying MozRank or PageRank by scaling the size of each node on the graph. Additional visualisation aid with extra parameters could be achieved through edge (link) colouring influenced by the number of outbound links on the page.

The above graph was produced by partial and combined Force Atlas 2 and Yifan Hu transformation with manual node colouring and colour-flow to outbound links.
When I was little I asked my grandfather how he makes his wooden sculptures. He told me that his job was to carve out the unneeded bits from a lump of wood. What’s left is the artwork. Google isn’t that different. In order to produce good results you have to carve out the bad ones. In this article I will demonstrate a little-known methodology used to evaluate quality and freshness of blogs, taking you beyond links, content quality, page layout and social signals. This post is dedicated to all the busy business owners who can’t find time to write very often but feel compelled to do so because ‘it’s good for SEO’.

Is fresh content good for rankings?

Let’s get it straight. Fresh content is good for where fresh content is useful. Search engines recognise this and will favour fresher and more up-to-date results ignoring the lack of slowly developing link signals for that URL. An old result can rank just as well as a fresh one for certain queries.

How often should I be posting on my blog?

Before I answer that let me point out the fact that search engines have plenty of content already. Most of the content produced today can be safely filtered out as low quality, duplicated, or as spam. In fact that’s why they unleashed Panda filter on us. You don’t want to be adding more stuff to that pile. Write when you feel like it or when you have something to say to the world. That’s what blogs are for. And here’s the ‘yes but’ part, and the reason I am writing this article...

Blog abandonment

An abandoned blog is a dead blog and Google knows that. Separating dead blogs from active blogs matters to Google. Why? It’s a quality signal, it’s a freshness signal and they may even use it to remove web spam from results. Yes. Really.

How is blog abandonment detected?

The first and most obvious way to check is to look at the timestamp of the last post. The problem with this method is that it does not factor the blog’s individuality. Each blog is driven by a different author or a team of authors and may exhibit wildly varying posting patterns. In addition to this, there is no such thing as a universal ‘blog expiry date’. Some blogs publish daily, weekly or monthly. From time to time you will run into a blog that only posts on a rare occasion covering a yearly group of events which may take place within a single week (for example yearly festivals).

The answer: User-centric metrics

Without generalising blogs, search engines may look at each individual blog and observe its activity history in order to ascertain its publishing model and detect posting patterns.

Academic research

Kerry Rodden (User Experience Research, Google) and Adam D. I. Kramer (Department of Psychology, University of Oregon) analysed roughly one million blogs tracking their habits including metrics such as total
number of posts, number of days between posts and the age of the blog (the difference between the first and the latest post published on the blog). They tried to identify "established" and "active" blogs and found that there is in fact a threshold point where they think that blog never made it and cannot be treated as established [[1]1]. Panda algorithm addition wiped-out blogs of this type a few years later.

What blogs are at risk?

Short burst blogs start with ten posts within their first week and show no activity after that. Visitors seeing the last post from 2001 will not be impressed and neither will be search engines. Short burst blogs like that are treated very differently than what Google considers 'established' blogs. Researchers found that by removing blogs with a life span of less than nine days overall quality of blogs in the collection increases.

Similarly, blogs with less than eleven posts in total were removed from the collection in order to allow researchers to focus on what they considered to be real blogs.

Research challenges

Established blogs demonstrate highly skewed mean times between posts and observations such as post frequency needs to be treated individually as it may represent different things with different blogs. What this means is that your blog may have a special little place at Google and frequency of posting and any reasonable deviations from the standard pattern may be known to them.

Blog Activity Perception

Once the blog has satisfied basic criteria of being established or substantial it can be assessed further to determine whether it’s active or not. What’s interesting here is that times between posts are observed on the blogger level (authorship signals) and not on the blog level. This is done to establish coherent observations for each blogger. This obviously does not affect single-user blogs.

Observing blogger’s activity it was found that nearly all posts will be less than three standard deviations above the mean. This is helpful to know when trying to determine if the blog has been ‘abandoned’ or if a blogger stopped posting there at some point and when.

Test case
In a test case Rodden and Kramer compare their method to 30-day metric and try to predict which blogs have not been established or have been abandoned. While observing active blogs they rejected 62% of blogs as not established and describe them as likely "fad" or "spam blogs". Abandoned blogs made up 6% of the lot with 32% of established blogs with recent posts. Inactive blogs were also observed and in that segment 2% of blogs were reclaimed from the "active" model with 31% showing abandonment and a massive 67% of never even being established.

Conclusion

Be consistent in your blogging activity, set realistic goals and follow through with your schedule. Wild deviations in posting habits could send the wrong signal to search engines. Worse yet, if you stop blogging altogether your blog may be labelled as abandoned. Surely that can’t be good for its rankings.

Keep in mind that this post illustrated only one temporal method of judging blog quality and eliminating spam.

Do not ignore community, user experience, social activity, content, links and technical aspects of your blog. Blogging goes beyond SEO and each great post you publish is another node in your network - ready to catch some more traffic.

References:

[1] [2] Applying a User-Centered Metric to Identify Active Blogs
Adam D. I. Kramer, Kerry Rodden
4.3 March

4.3.1 Refining Search Engine Queries: Clustering and User Intent (2012-03-07 10:15)
- admin

Understanding the way that search engines determine which pages to display in response to a user query can help to enlighten business owners and end users regarding the rankings and page views they receive from these search engines. A recent study conducted by Stanford University in conjunction with Google Inc [1]. explored the various ways that queries can be refined and related more closely to the intention of the end user. A variety of methods were examined and evaluated in order to determine the most effective ways to handle related queries and to provide the user with the information they need more effectively and rapidly.

Diversity

Presenting a number of diverse results can provide the end user with a wide range of choices. This method assumes that the user will scan through these widely variant results to find the one that best meets his or
her needs. However, in many cases the number of variants may be overwhelming for the end user and may not produce the correct results even when the most relevant topics are included in the results.

Clustering

A more advanced method of ordering search results groups related pages together. For example, pages about apples would be clustered together in one group, while Apple computer pages would be included in a separate cluster on the results page. This provides the benefit of choice to the end user while still allowing a larger sampling of relevant pages to be displayed for each search engine query. Most advanced methods of ordering search engine results use some form of clustering to display those results to the user.

Multi-session methods

Sophisticated search engine query algorithms can take into consideration the previous search activities of the IP address or user in order to provide the most relevant results. By considering recent searches when returning query results, search engines can provide faster, more accurate service for their end users; this is especially true when those results are clustered into relevant groupings.

Putting it all together

Clustered search engine results are preferred by users, as they provide an optimal degree of diversity while maintaining a manageable list of potential results. By refining the search engine algorithm’s clustering method with information about recent searches and deriving information about the user’s intentions from those searches, query results can be tailored to the specific needs of the end user. This allows the most relevant and immediately necessary results to be ranked more highly and displayed more prominently than those that are less likely to render the needed information to the end user.

By clustering query results by general topic, search engine providers can better serve their end users and provide more accurate page rankings and query responses for all parties involved.

References

The next stage in Google’s product evolution is undoubtedly unification. Google’s many projects and shatterlings spread over a decade of development in a frantic search for the next big thing are coming back home, empty handed. Now they’re merging into a single universal product, bit by bit every piece of knowledge and experience collected so far is being used and integrated into a bigger, better and faster Google. Everything else will perish.

This is the path we’re headed down – a single unified, ‘beautiful’ product across everything. If you don’t get that, then you should probably work somewhere else. Larry Page, Google

This bold statement was directed at Google employees during a staff event following the Search Plus Your World launch. It’s a clear evidence that the information giant is not kidding and are raising the bar for themselves, and their competition.

In 2012 we are to see more product and service canonicalisation at Google. [1] Old and tired projects will be chopped up to bits and the best parts will be integrated into something new, similar products will be merged together and Google’s social network will continue to grow and further integrate with everything else.

***

Google isn’t going to wait for a product to be fully baked before releasing because if they did it would stifle innovation (thats why it’s the Google+ project not product)

Is this the end of Google’s innovation?

Certainly not. In fact it has been mentioned many times so far that Google’s new policy towards product development is to more or less send it “out there” and field test by collecting data and user feedback in order to rapidly improve result quality, UX and functionality. We’re already seeing examples of this attitude toward product launch in a rapidly evolving Google+ platform and Google Webmaster Forum re-design.
What does Google ‘own’ so far?

Here are a few key things:

- Search: images, videos, academic, news, mobile, books, shopping...etc
- Local: Maps, reviews, business listings
- Email: Gmail, chat, video/audio
- Tools: Translations, directions, conversions, insights
- Office: Google docs, storage, calendar
- Enterprise Solutions
- Social Network: Google+
- Projects & Acquisitions: YouTube, Blogger...etc
- Browser Market: Chrome
- Operating System: Android
- Hardware
- IP: Research, patents, methodologies, algorithms
- Philanthropy, charity, PR, non profit, science and education relations

All these projects and products require different teams in different countries and coordination of activities takes up a lot of time and energy. And of course there’s money. Many take Google for granted and forget they are a business.

Last Year’s Announcements

In 6 months Google+ is going to look like a different product as it integrates with other Google products. Google made many hints and promises of new and exciting things to come and many have already been released. To illustrate I quote [2] Carter Gibson’s notes from CrushIQ in November 2011.

Early signs of this announcement were visible already in December 2011 with January and February 2012 rolling out some major changes including the release of Search Plus Your World (also known as SPYW, or simply Search Plus).

Google’s New and Unified Privacy Policy

“One policy, one Google experience”


On March 1, 2012 Google released a privacy policy that is not new, but actually consolidates all the policies from across different Google-owned platforms like YouTube, Gmail, Google+. The intention of unifying their
policies was to make sharing information across Google with other users much easier and error-free. All information that is collected and stored by the company is held on secure servers and limits access to only employees with proper authorisation.

Google collects information from users in various ways. This includes information provided when signing up for services, information gathered from the use of services including device, location, log information. The personal data collected is used to create a more personalised and consistent experience for users. This information is shared with a variety of users for different reasons including legal issues and companies that a user consents to sharing with. Google will also aggregate some of the personal data into non-personally identifiable groups and use for other reasons.

Users of Google services have the option and ability to edit personal data that is collected and limit the ways the data is used. Each user is also able to edit any profile information that is publicly shared and control how particular people or groups of people view one’s profile.

Google Play

[blockquote type="blockquote_line" align="right"] Now your favorite music, books, movies, apps, and games are all in one place that’s accessible from the Web and any Android device. Discover, buy and share like never before. [4]play.google.com[/blockquote]

The month of March has seen Google usher in a unified privacy policy for the majority of its products. The consolidation continued last week as Google introduced its new Google Play brand. Google Play is a new one-stop destination for users to get content. To be clear, Google Play is essentially just a re-branding, although it does present some new features. With the launch of Google Play, Google has done away with a few separate services. Gone are Google Music, Google Books, and the Android Market. Instead, all of the content that used to be provided under these different brands is now available under the Google Play label. Users can now log on and choose from more than 450,000 apps for Android devices, as well as thousands of movies and millions of songs and books.
The main point of introducing a new service was to simplify things, and Play certainly does make things a lot easier for regular users of Google products. Prior to the new central service, Google Books users couldn’t access their movies from the same place, and users of the Google Music app had to go to the Android Market to purchase songs. Everything has become a lot more intuitive under the new label.

Cloud services are a large part of Google Play’s benefits. Users can purchase a book from Google Play and read it online or on an Android device. Music and movies purchased on an Android device, computer or Google TV unit can be listened to or watched on any other device. All a person needs to do is log in with his or her Google account. Everything is connected together.

Perhaps most appealing of all, the entire new service is free. There’s no need to pay for cloud accessibility. That means users of devices with limited storage don’t have to worry about backing up music they’ve bought should they need to delete it to make room for new content. Google Play users can even add upload up to 20,000 songs from their personal library and access it from any Android device or web browser.

Overall, it looks like a smart move by Google so far. Users have more options and less hoops to jump through to find the content that they want. Free previews of movies and books are also available to help make the purchasing decision. With so much functionality packed into one place, the Apple’s iTunes ecosystem looks to have a legitimate competitor in Google Play.

Leaner and More Focused Google

Resolutions can be hard, and changing products that people love is hard too. But we’re excited to focus on creating a beautifully simple, intuitive user experience across Google - an experience that will change the lives of millions of people. Dave Girouard, Google

Although this is partially true, we must factor in the recent product trim announced by Google. As part of their new year resolution for 2012 Google announced that they will be looking at products which have
elements same or similar to other products, those who have not lived up to their potential or simply cannot be integrated with other products and features within Google.

They explain that this move is in order to gain more focus and time for main products and make them work well. Not a bad idea.

Here are a few products which are now officially phased out:

- Google Message Continuity (GMC)
- Google Sky Map
- Needlebase
- Picnik
- Social Graph API
- Urchin

Support, Help & Communication

Webmaster Tools

In 2012 Google’s communication with webmasters and website owners is at an all time high, and not as a sudden surge, but a rather well-planned and gradually executed mission. In 2011 webmasters have seen more security alerts and a new variety of messages for troubled pages and domains. This year brings new addition to an ever growing range of emails being sent out by Google.

Here are some alert examples:

- [5] Outdated CMS
- [6] Unnatural links
- [7] Low quality pages
- [8] Traffic change

Webmaster Help Forums

In addition to [9] Google Webmaster Tools, Google encourages discussion on their [10] Webmaster Help Forums, which have among many other sections, undergone a [11] total design and functionality re-vamp. The forums currently feel slightly out of place with some functionality, but Google assures us they’re on the ball and more improvements are coming.
Google+ Hangouts

Google also actively used [12]Hangouts to communicate with small groups of webmasters (up to ten at any time) in multiple locations, time zones and languages (English, German, Hungarian...etc). We've even seen [13]Matt Cutts join in and answer questions in a January hangout.

Public Criticism

SPYW: Social or Creepy?

[blockquote type="blockquote _line" align="right"]Google policy is to get right up to the creepy line and not cross it. Eric Schmidt, Google[/blockquote]Google's prime focus today is the place where link graph meets social graph. Some early reports show that users find personalised search results a little bit too creepy and prefer to have the old results back. How reliable this data is will be revealed through further surveys, but ultimately by user reactions. One way of semi-gauging personalised search adoption would be by monitoring of the (not provided) search referral keyword in Google Analytics.

Non-Inclusive and Proprietary

Many in the search community feel that Google's attitude change towards [14]third party platforms has changed dramatically - and in the wrong way. Google has been systematically avoiding any integration of third party social media platforms and their functionality in any of its new product releases.
We asked them about this and the answer we received was basically that Google does not feel comfortable investing time and money in building up products and infrastructure to support platforms which may change or turn off access for Google any time they want. We were also reminded of the recent Twitter integration fiasco and the loss of firehose access.

In one way this makes perfect business sense. But when you think about it, what has ever stopped Google from accessing information from different blogging platforms, forums, chats, office documents and other exotic data? Only robots.txt and similar protocols.

In a recent [15]interview with TechCrunch, Vic Gundotra (Google) makes another less than convincing claim on why Google+ won’t let 3rd-party apps post and among other things mentions that “Your Stream Could Easily Be Overwhelmed”.

One other example of this behaviour is lack of support for Vimeo video embedding on Google+ stream. Are we expecting Vimeo to suddenly change their embed code and break everything on Google+? Certainly not. Facebook has done it for a while now so it must not be something you would describe as a great technical challenge.

The Future

Sometimes we need a reminder that Google is a business. They grow, change, make mistakes, learn and improve. I run a company myself and know that my business is only as good as my people. I dare to say that the future looks good for both Google and its users and can claim this on the basis of my personal experience and encounters with many great Googlers who work hard at making their search engine better. This includes the Search Quality Team (including Web Spam Team), Webmaster Help Team, Community Managers, Researchers, Product Developers, UX Specialists and many other departments.

IFRAME: [16]http://www.youtube.com/embed/JtRJXnXgE-A?rel=0

Your Thoughts?

What do you think will happen next in the world of Google? Will their direction remain the same or do you
see a shift take place in the future? Please leave your comments below or on [17]Google+.

4. https://play.google.com/about/
17. https://plus.google.com/114074532743058808065/posts/YBQjh8LqHfm

4.3.3 2012 Brings 3 Key Improvements in Google’s Algorithm (2012-03-17 13:36) - admin

Google is determined to level the playing field, reward good content and punish the cheaters. How will they achieve this?

There is plenty going on at Google at the moment but let’s highlight three key elements relevant to this article, even though they may not be as obvious at first:

1. Improved spam detection
2. Knowledge graph
3. Social Graph

Improved Spam Detection

Web spam and aggressive SEO tend to have a domino effect due to lack of ‘justice’ and equal treatment of all websites by Google. If one webmaster gets away with cheating the others will follow. In 2010 we wrote about this in detail: [1]Downward Spiral of Questionable SEO Practices and Why White Hat SEO Struggles to Survive.
This year SXSW featured a panel named "Dear Google & Bing: Help Me Rank Better!" with Danny Sullivan (Search Engine Land), Duane Forrester (Bing) and Matt Cutts (Google).

During the panel discussion Matt Cutts announces an upcoming change within Google’s algorithm. The change is all about catching overly aggressive SEO signals and taking action against affected pages:

"We are trying to level the playing field a bit. All those people doing, for lack of a better word, over optimization or overly SEO – versus those making great content and great site. We are trying to make GoogleBot smarter, make our relevance better, and we are also looking for those who abuse it, like too many keywords on a page, or exchange way too many links or go well beyond what you normally expect.

You can download the recording of the session as an mp3.

What elements might Google be after?

1. Keyword stuffing
2. Thin content and SEO themed pages (e.g. entire page dedicated to car rental + postcode).
3. Automated link exchanges
4. Open invitations to game the system (on-site, via email, posted externally)
5. Link buying
6. Anchor text over-optimisation (excessive percentage of commercial terms in backlink profile)

7. Aggregating and scraping duplicate content for SEO reasons (inserting own links in press releases)

8. Unnatural links (spam, article submission, blog networks)

9. Microsites and own domains to gain extra popularity in search engines

10. ”Clever” and ”creative” but sneaky tactics (including anchor text based widget distribution)

It’s a bit surprising they are announcing this considering that they did take action against many of these schemes already. Perhaps this time they have an algorithm which can determine over-optimisation with a greater degree of certainty.

As usually we expect all borderline cases to receive warning emails prior to any filters and penalties being applied. Best thing to do now is to scan [4]Google Webmaster Forums in the next 2-4 weeks.

Knowledge Graph

Google engineers are not satisfied with the quality of their search algorithm and admit that their search engine is not that good at truly understanding the user query.

Where they want to be is in a place where user asks a question and gets a solid answer instead of many choices. This is of course applicable in only certain situations and some elements of that are already visible in search: [5]Conversion, [6]mathematics, [7]schedules, [8]technical data and [9]facts.
The Wall Street Journal, published an article titled "Google Gives Search a Refresh" featuring Amit Singhal (Google) giving us a glimpse of a direction the world’s biggest search engine is taking. Prior to that interview Singhal was also featured in Mashable where he talked about Google’s "knowledge graph" and the impact it could have on search.

Rather than re-inventing the wheel Google found an interesting acquisition target "Freebase" and knew what to do with all that data. Freebase is basically a third addition to the "graph" entity at Google:
This acquisition may seem like a quick hack, but Google assures us that [13]they’re taking the linked data they have much further and using the Freebase knowledge graph as the initial seed to create much larger knowledge collection of interconnected entities and their various attributes. Google [14]grew Freebase database from 12 to 200 million entities and continues to invest heavily into its growth and development.

Social Graph

Google has [15]unified and merged many of its services and as part of that we can expect to see their growing social graph to start making greater impact on search results.
This change will not only manifest through explicit means such as Search Plus Your World and Google+ search but will find its way into classic search results where it will represent a meaningful signal and a ‘validator’ of link and knowledge graph blend. Part of this mix will naturally be authorship signals and their own internal "AuthorRank" for bloggers, authors and journalists, or "PersonRank" in context of typical social interactions.

The key to social graph is of course it’s structure and our connectivity with other nodes (people) within the social graph. Google calculates connectivity on two levels: Implicit & Explicit. This means that they use both our own ‘declared’ connections and the ones we form unknowingly through social interaction, sharing, +1’ing and commenting on Google+.

Your Thoughts

We would love to hear what you think will be the biggest changes in Google in 2012. Please leave your ideas in the comments below or on Google+ discussion about this article.

2. [http://schedule.sxsw.com/2012/events/event_IAP11742](http://schedule.sxsw.com/2012/events/event_IAP11742)
6. [https://www.google.com.au/search?q=(sqrt(cos(x))*cos(400*x)+sqrt(abs(x))-0.4)*(4-x*x)^0.1&pws=0](https://www.google.com.au/search?q=(sqrt(cos(x))*cos(400*x)+sqrt(abs(x))-0.4)*(4-x*x)^0.1&pws=0)
4.4 April

4.4.1 Search Quality: The Link Graph Theory (2012-04-07 18:52) - admin

Theory behind evaluation of search result quality through algorithmic treatment of the web index. Read on and discover some of the fundamental forces behind search engine rankings.

Part 1 of the [1]Search Quality Series by Dejan SEO

Introduction

There are different types of graphs on the web, useful to search engines in processing and evaluation of relevance and reputation of pages in their index. Here are examples of graphs currently being utilised by search engines:

1. Link Graph: Nodes (pages, documents) connected by directed edges (links).
2. Co-Citation Graph: Nodes connected by undirected edges (A & B are connected if C links to both)
3. Social Graph: Explicit and implicit connections between individuals on the web.
4. Knowledge Graph: A collection of interconnected canonical entities and their attributes
Link graph aids in the ranking of documents while co-citation makes it easier to find relationships and categorise documents. Social and knowledge graphs are relatively new additions with ability to answer our questions directly or personalise our search experience.

Connectivity-Based Ranking

Query-Independent

Google converts its entire index into a graph in order to (among other things) calculate connectivity-based ranking for that collection. The first mode of connectivity-based ranking system is a query-independent ranking model based on document connectivity alone. This is where PageRank algorithm comes in handy. In a nutshell, PageRank is the value of probability of users staying or leaving the document. PageRank alone, however, is incapable of producing relevant results and simply reflects the perceived importance of the page on the web.

A document which points to many others might be a good hub, and a document that many documents point to might be a good authority. Recursively, a document that points to many good authorities might be an even better hub, and similarly a document pointed to by many good hubs might be an even better authority. , Monika Henzinger, Google Inc.

Query-Dependent

The second mode of connectivity-based ranking is query-dependent and capable of slicing the link graph into a relevant sub-set which can be re-ordered and re-ranked in a way more meaningful to a given search query. In addition to a standard indegree model (inbound links) two additional factors are introduced to improve relevance of returned documents:

1. Hub score
2. Authority score

Both are part of the HITS algorithm which stipulates that hubs are nodes which link to pages relevant to the search query and authority nodes are the pages expected to have relevant content.

The concept of query-dependent / connectivity-based algorithm in action is illustrated here:
The above image represents a micro-web of not more than 10,000 domains and the ranking system is calculated on a query-independent principle. In the second example the algorithm identifies only those nodes which are found to be substantially related to the user query:
This allows for removal of query-irrelevant nodes and calculation of reputation only within query-specific context.

Simplified Visual Examples

In the following graph we’re seeing authority in action:
Note that authority works on an egalitarian level and exhibits disregard for inbound link qualities. Another quantitative node analysis is hub detection:
Finally, let’s see how PageRank solves the problem of equal count by introducing a more elegant way of ascertaining ranking value of documents in our collection:
A combination of all three will produce the best results with first two identifying topical relevance and the last one for sorting purposes. There are in total five key elements to Google’s algorithmic link analysis and processing:

- In-degree (inbound links)
- Out-degree (outbound links)
- HITS authority score \([8\text{][2]}\) (quantity of relevant inbound links in a sub-set of the index)
- HITS hub score (quantity of relevant outbound links in a sub-set of the index)
- PageRank (random surfer model)
Together the above mentioned methods form a fundamental link analysis unit which is further adjusted and enriched with additional treatment of data and various, more subtle refinements of data.

Top-Level Spam Detection

Pictured below is a multi-mode visualisation of an artificial blog network consisting of 4762 domains and 9849 connections between domains. Note that we collapsed each domain into a single node in order to simplify the visualisation. The practice of regarding all nodes within a single domain as a single node is not uncommon and has application as one of the steps in evaluation of query-dependent rankings.

The observed network is real (sent to us by a genuine spammer offering to buy links). Numeric values attached to each node in the graph represent the internally calculated PageRank score and have been used to scale and colour the nodes accordingly. Highlighted in green are the top link recipients within the collection.
ranking algorithm manipulation. Same principles for detection of relevance and topical similarity can be applied to look for abnormal patterns in the linking structures of any sub-set of the link graph. For example, here are the top perceived hubs within this network:

Artificial linking patterns are evident even after rudimentary visual examination, but further statistics can be extrapolated through several statistical parameters including:

1. Network Overview:
   1. Average Degree (weighted / non-weighted)
   2. Network Diameter
   3. Graph Density
4. HITS
5. Modularity
6. PageRank
7. Number of weekly and strongly connected components

Node Overview:
1. Clustering Coefficient (including average)
2. Eigenvector Centrality

Edge Overview:
1. Average Path Length
2. Neighbourhood Overlap / Embeddedness

All of the above parameters are contained and available within Gephi platform [[11]3]. Here is an example of statistical analysis of the blog network:

![Eigenvector Centrality Distribution](image)

Note that in Google’s algorithm, PageRank plays the role of the dominant eigenvector of the probability matrix as part of the 'random surfer model' [[12]4].

Consider the number of options available for spam detection, all this without any aid of additional signals and metrics available including, anchor text, content and many other technical patterns. By introducing extra elements in spam analysis Google can predict with great certainty if a page exhibits unnatural structure or linking patterns.
Practical Takeaways

1) Linking out to authoritative content is a property of a hub. Hubs are important to search engines to define topic and categorise pages.

2) Search engines can infer topical relation between the two sites without an explicit link by using co-citation. Query relevant hub A links to both B and C so they must be also relevant to the search query. A virtual link between the two is formed.

3) Avoid artificial link structures as they are discoverable on many levels.

References:

4.4.2 Link Spam Detection (2012-04-12 15:38) - admin

In this article we describe quantitative metric-based link spam detection in a collection of analysed domains. Accuracy and co-relation ranges from 60 to 90 percent with adjustment of different quality parameters. Our technique is still considered to be ‘work in progress’ and may not be fit for accurate automated action. Instead we offer a rudimentary, top-level analysis method, suitable for first round flagging and creation of spam alerts within large collections of domains and massive backlink profiles.

In the above image we show the case of strong co-relation with the metrics on the right and the manually flagged spam domains showing in red on the left.

Metrics & Calculation
Although not all metrics proved to be useful in our analysis we will list all included in our spreadsheet (from left to right starting with column A and ending with W):

1. Domain
2. Backlinks (BL)
3. Maximum Backlink PageRank (MaxPR)
4. PageRank Sum (tPR)
5. Unique PageRank Sum (uPR)
6. Unique Domain Backlink Sum (udBL)
7. D-Factor (D=uPR/udPR)
8. Average D-Factor Difference (avgD=D-(D1+Dn.../n)
9. Unique Government Domain Links (uGOV)
10. Unique Educational Domain Links (uEDU)
11. Trusted D-Factor (tD=uGOV/uEDU)
12. Average Trusted D-Factor (avgtD=tD-(tD1+tDn.../n)
13. Advanced Trust (aT=avgtD/avgD)
14. Manual Trust Value (mT=Manually entered for benchmarking purposes)
15. Domain Length (L)
16. Formula Variants:

1. \[ S1(n10)=((F2*E2)/100+(J2*10)+(I2*10))*N2 \]
2. \[ S2(n10v)=((F2*E2)/W2+(J2*10)+(I2*10))*N2 \]
3. \[ S3(n100v)=((F2*E2)/(W2)+(J2*100)+(I2*100 ))*N2 \]
4. \[ S4(m)=((F2*E2)/(W2)+(J2*100)+(I2*100))\times M2 \]
5. \[ S5(h)=((F2*E2)/(W2)+(J2*100)+(I2*100))\times H2 \]
6. \[ S6(l)=((F2*E2)/(W2)+(J2*100)+(I2*100))\times L2 \]
7. \[ Sx=(E2*F2)*H2 \]

Booster (Arbitrary value of 1000, used for smoothing the detection highlighting gradient and can be adjusted at free will)

Preliminary Results
The most accurate results were achieved through following formulas:

- S4(m)
Simplified formula (Sx) is considered to be a borderline case, however it does offer a massively simplified method of link spam detection which proves to be useful even with considerably degraded result accuracy. The logic behind the simplified formula is as follows: \( (uPR \times udBL) \times D \).

Likewise aT metric gives reasonable accuracy measure on its own. In the table below the first two rows are accurate guesses (aT values of -0.70 and -0.74) with the other three being false positives. We therefore predict that the aT threshold value in this particular collection is above -0.60, which in our collection of observed domains accurately separated the spam from genuine websites.

As a reminder aT is a sum of Average D-Factor and Average Trusted D-Factor, this means that introduction of edu and gov domains does indeed hint at quality, though we give .gov more weight after observing more than 100 flagged results. Even though .gov domains seem to be better moderated and harder to infiltrate by spammers they are not entirely immune to manipulation (e.g. spam of the public log files and statistics). This is where qualitative analysis comes in, however this is outside of the scope of our study.

In the table above, last four columns illustrate our more complex formulas which tend to place a wider gap between the domains with organic and inorganic links.

Spreadsheet Access & Comments

To request access to our spreadsheet or make a comment or suggestion please visit the [Google+ post](https://plus.google.com/u/0/114074532743068808065/posts/VpAzzfYrKV2) for this article.

References:

References and research which inspired our work include:

- [3] Quality-Biased Ranking of Web Documents
- [5] Spam Behavior Analysis and Detection in User-Generated Content on Social Networks

2. [https://plus.google.com/u/0/114074532743068808065/posts/VpAzzfYrKV2](https://plus.google.com/u/0/114074532743068808065/posts/VpAzzfYrKV2)

### 4.4.3 Selective Sharing in Google+ (2012-04-20 00:49) - admin

Review of the first empirical study of Google+ platform in the context of selective sharing.

Real-World Sharing & Social Networks

Sharing information is a crucial part of human interaction and communication, however in the offline world we manage the information we disclose in a natural and effortless way. We decide what and when to share
with different facets of our social circles.

Social networks however have for a long time practiced the ‘all or nothing’ principle where shared information may reach wider audience than users may prefer. One paradox tied to social networks is the phenomenon of ‘over-sharing’, even by users who are aware of its potentially negative consequences.

Academic Research

In a paper called “Talking in Circles: Selective Sharing in Google+” [1], a team of researchers from Stanford University and Google examine the process of selective sharing within Google’s new social media platform.

Research activities outlined in the paper included an elaborate field trial and analysis of sharing patterns extracted from a large-scale dataset. Some of the observed behaviours include sharing with public, direct users and limited Circles.

The initial study looks at ways ‘pro users’ (or early adopters) interact and share information on Google+ assuming that their sharing patterns may be more instinctive and intuitive due to platform familiarity and general affinity towards new technologies.

The three main characteristics of Google+ highlighted as particularly helpful in social interaction online were: Circles (Selective disclosure of information), Extended Circles (Not covered in the paper), +Mentions (User referencing). Interesting similarities and parallels were drawn with other platforms for comparison reasons.

Data Collection & Analysis

The first step in quantitative log analysis was to examine the user-to-user sharing patterns. For this purpose 100,000 active users were randomly selected and their direct shares between 20th and 20th July 2011 were aggregated and analysed with user IDs being hashed for privacy reasons.

Sharing Style Breakdown

The graph below represents a breakdown by sharing types including public shares, selective sharing and targeted user sharing. The percentage values are indicating the breakdown of users who chose to share via each method at least once during the observed period and the percentage of items shared by same qualified users.
An interesting aspect of the analysis is that it included the most commonly used names for custom created circles, something that is otherwise not available for analysis due to privacy reasons.

The analysis took into consideration the top most common Circle names, and special attention was paid to top 50 occurrences. Other language variants were translated and included in the English pool of common Circles.

The top level observation was that the circles break down into two top-level categories which the paper labels as:

- Life Facets (work, school, technology, camping)
- Tie Strength (friends, family, co-workers, acquaintances)
  - Strong (family)
  - Weak (acquaintances)

In the graph below we can see the top 10, 100 and 1000 most common name categories broken down by two main facets and tie strength:

The research conducted hints at a third major Circle classification: topical interest group. For example a user may have Circles which include: Techno Music, Oil Painting, eBay Sellers and Cat Owners.

What is Being Shared?
Google+ users mainly share URLs, followed by photos and videos with location being the least shared type.

Why Do People Share Stuff?
Inherent Content Value
Nearly 60% of respondents explain that one of the major driving factors for sharing is the inherent value of the content.

Value to Others
Another common reason is the perception of value to others. Useful and informative content gets shared.

Emotion, Interest & Curiosity
It’s no surprise that shared content was described as interesting and cool. Content that’s positive in nature also attracts shares.

Entertainment
Around 15% of respondents highlight funny and silly content as something this share for its humour and entertainment value.

It’s About Me!
A quarter of users in the interview admitted to sharing content about them personally whether it is an update or a call for peer-review of opinion or a personal view.

Eager to Discuss & Participate
Google+ users seek help from their connections and offer help in return. Desire to participate in a debate or discussion on a topic was recorded in about 17% percent of surveyed users.

Supporters & Advocates
Slightly less common but certainly a significant fraction of nearly 7% of included Google+ users wish to raise awareness to a certain cause or entity.

Sharing Style & Frequency
The following graph summarises the self-reported frequency and type of sharing among selected Google+ users:

Who to Share With?
More than a half of all posts from observed users were shared publicly. The main identified elements and considerations in the context of selective sharing were:

- Privacy
- Relevance
- Social Norms
- Distribution
Circle Management

Four main circle-management strategies were identified:

- Inner Circle (tell-all, close friends, family)
- Structured Groups (presence of hierarchy)
- Interest Groups (topical relevance)
- Catch-All

Benefits of this study include potential advancements in contact management which may benefit from better breakdown logic reflecting user’s life facets, tie strengths and topical interests.

References:


Note: This is a review of a draft-stage academic paper. The camera-ready version will be presented at the proceedings of the ACM Conference on Human Factors in Computing Systems (CHI ’12), 2012

1. file://localhost/mnt/ext/blogbooker/tmp/dkn4b9xs/dkn4b9xs-body.tex.lynx.html#one

4.4.4 Why Google Went Social (2012-04-24 15:25) - admin

Let’s take a look at some of the confirmed ranking factors and find out the reasons behind Google’s social focus.
Our first search quality article discusses the link graph theory, co-citation, social graph, knowledge graph and top-level spam detection techniques. This time we go slightly further into specific search ranking factors and key events which lead to Google’s change of policy towards 3rd party providers and development of its own social platform.

Everything we discuss and quote in this article is detailed in the following paper: Indexing The World Wide Web: The Journey So Far by Abhishek Das and Ankit Jain, Next Generation Search Engines: Advanced Models for Information Retrieval, 2011. The paper covers much of Google’s fascinating crawling and indexing technology in great detail, however in this article we will only cover the items of interest to the SEO and internet marketing community.

On-Page Ranking Signals

Let’s see what Google engineers have to say about on-site ranking factors:

"First note that page structures, such as titles and headings, and url depth play a major role. Next we see that most terms occur close to each other in the results, highlighting the need for term positions or phrases during indexing. Also important, though not clear from the figure, is the respective position of terms on pages; users prefer pages that contain terms higher up in the page.

Other than these, search engines also learn from patterns across the web and analyze pages for undesirable properties, such as presence of offensive terms, lots of outgoing links, or even bad sentence or page-structures.

The diversity and size of the web also enables systems to determine statistical features such as the average length of a good sentence, ratio of number of outgoing links to number of words on page, ratio of visible keywords to those not visible (meta tags or alt text), etc."

Brief Interpretation

Here’s a quick overview of the key ranking factors outlined in the paragraph above.

Strong Signals

- Title, H
- URL depth
- Term proximity in a phrase
- Term position on a page
- Order/sequence of phrases

Negative

- Offensive language
- Many outgoing links (low quality?)
- Bad grammar / spelling / style
• Bad page structure

Statistical Optimum

• Average length of a sentence
• Number of outgoing links VS page word count
• Visible VS meta content ratio

Off-Page Ranking Signals

Here are the two key sections of the paper which talk about off-page ranking signals:

"...off-page signals have increasingly proved to be the difference between a good search engine and a not-so-good one. They allow search engines to determine what other pages say about a given page (anchor text) and whether the linking page itself is reputable ([4]PageRank or [5]HITS). [...]"

The final ranking is thus a blend of static a priori ordering that indicates if a page is relevant to queries in general, and a dynamic score which represents the probability that a page is relevant to the current query. [...]"[6]HITS, as described earlier, scores pages as both hubs and authorities, where a good hub is one that links to many good authorities, and a good authority is one that is linked from many good hubs. Essentially, hubs are useful information aggregations and provide broad categorization of topics, while authorities provide detailed information on a narrower facet of a topic. [...]"

However, instead of pre-computing the hub and authority scores at indexing time, each page is assigned a query specific score at serving time."
4.4. April

BlogBook

Brief Interpretation

Off-page signals, [7]PageRank, [8]HITS ([9]hubs & authorities), anchor text and so on, combine as a layer on top of query matching on-site. Google looks at and trusts pages which only link to authoritative sites and ranks sites linked from trusted hubs. HITS related scoring is assigned at the query time and for a particular search segment, not like PageRank, which is pre-calculated and simply assigned to enhance sorting/scoring of results.

Anecdote: "...webmasters cannot control which sites link to their sites, but they can control which sites they link out to. For this reason, links into a site cannot harm the site..."

Future Direction: Real Time Data & Search

In this section we highlight key areas of the paper which talk about Twitter, Facebook and hint at reasons behind creation of Google’s own social search layer and social media platform.

"In 140 characters, users can describe where they are, publicize a link to an article they like, or share a fleeting thought. From a search engine perspective, this information is extremely valuable, but as different projects have shown over the last few years, this information has extra importance if it is mined and presented to search engine users in real time."

Twitter Firehose

Social graph is not a big deal for Google as they already have the algorithm and infrastructure to deal with linked data. People are, in fact, well-linked and this is something search engines look at. At the time, access to the firehose was taken for granted, and appreciated:

"In order to build a real time system, the first prerequisite is access to the raw micropost data. Luckily for the community, Twitter has been extremely good about providing this at affordable prices to anyone who has requested it."

Twitter Firehose

Social Ranking Signals

Here are the key steps in dealing with [10]real-time / [11]social data:
1. Create a Social Graph: UserRank, UserTopicRank
2. Extract and Index the Links
3. Real-Time Related Topic Handling
4. Sentiment Analysis (theoretical only)

Social graph in Twitter for example is understood through analysis of users’ connections (following, followers) and topics they are interested in (what we tweet about, what people we follow tweet about). Number of followers is another useful and measurable metric Google uses to determine the UserRank and even detect topic leaders, users influential within a certain niche (UserTopicRank).

Links within tweets are crawled and indexed. Standard link metrics such as anchor text and PageRank are replaced by the link context and UserRank respectively.

Finally, classic web index domain quality parameters can be added to social layer to enhance the results. This is a great example of how different index types at Google work well together to produce seamless search experience.

Thanks to real-time data Google can define and cluster fresh topics and serve the right results at the right time (e.g. election, football game) and default back to ‘stale’ results when the popularity peak diminishes to usual levels.

Facebook & Bing Social Layer

The following few paragraphs truly reflect some of the key events which compelled Google into abolishing partnerships and alliances with third party providers and trusting only its own properties:

"Over the last few years, Facebook has become the leader in social networking with over 500M users [Zuckerberg, 2010]. Facebook users post a wealth of information on the network that can be used to define their online personality. Through static information such as book and movie interests, and dynamic information such as user locations (Facebook Places), status updates and wall posts, a system can learn user preferences.

Another feature of significant value is the social circle of a Facebook user, e.g. posts of a user’s friends, and of the friends’ friends. From a search engine’s perspective, learning a user’s social interactions can greatly
help in personalizing the results for him or her.

Facebook has done two things that are impacting the world of search. First, in September 2009, they opened up the data to any third party service as long as their user authenticate themselves using Facebook Connect [Zuckerberg, 2008]. Second, as of September 2010, Facebook has started returning web search results based on the recommendations of those friends who are within two degrees of the user [...]

In late 2009, Cuil launched a product called Facebook Results [Talbot, 2009], whereby they indexed an authenticated user’s, as well as his or her friends’, wall posts, comments and interests. Noting that the average user only has 300-400 friends with similar preferences and outlooks on the world, one of the first discoveries that Cuil made was the fact that this data was extremely sparse.

This implied that there were very few queries for which they could find useful social results. They overcame this limitation by extracting related query terms, which allowed for additional social results.”

Brief Interpretation

To Google it was obvious that they lack the platform and technology to collect the wealth of data in the ever-growing social media sphere. A few attempts and negotiations were made in order to capture the slice of the prolific social user content pipeline; Google knew what they needed and what they had to do:

1. Facebook opened data via authentication
2. Web search results based on user recommendations
3. Static Information (Interests)
4. Dynamic Information (location, status/wall)
5. Learning User Preferences
6. Social Circles
7. Personalisation of Results

Note: Small circle personalised results too shallow, related query extending useful to enrich choice.

The Last Straw

"In the future, web search engines can use such a signal to determine authority of social data. In October 2010, Bing and Facebook announced the Bing Social Layer."


Google’s Reaction
Series of events took place during this time including strategic acquisitions, change in management, policy and attitude shift towards 3rd party providers, product trim and focus on getting what Facebook and Twitter have. Academic research on social networks, user experience expertise and already developed infrastructure were adapted to handle big-data processing and algorithmic treatment of linked data.

Google announces SPYW, Google+ is launched. Top level management decides Google is now one unified product and utilises enormous search market share, brand, product range and influence to propel the success of the new social platform. Rapid deployment policy continues and Google’s products including their social media platform continue to evolve and change, while authorship becomes more significant and easier to implement. Attempts of public criticism are largely ignored or downplayed.

The search giant has moved on and even though Twitter and Facebook are being looked at and factored in, they will never again be in Google’s circle of trust. Ever?

Stay tuned.

Discussion

Opinions, comments, questions & feedback on Google+

References:


4.4. April

BlogBook

9019.pdf

29. https://plus.google.com/authorship
32. https://plus.google.com/111588754935244257268/posts/N9kx71RQR1U

4.4.5 Enterprise SEO: 2012 and Beyond (2012-04-29 23:17) - admin

Learn about the key elements behind our most successful enterprise SEO campaigns.
Enterprise SEO brings a whole new set of challenges on top of usual set of best optimisation practices and one of the most momentous obstacles with large-scale campaigns is the sheer scale of things.

The General Rule of Scalability
Avoid manual work and find ways to automate anything repetitive or large in scale.

Scalability is the key ingredient which spans across various facets of big campaigns including strategy design, research activities, link building, technical assessment and implementation. For example in multi-dimensional product/service websites, focusing on a single vertical can be useful as a matter of testing a new method or coming up with a proof of concept. Usual SEO workflow structure will simply not be scalable enough to execute across thousands of pages and keywords.

So whatever you do, do not get stuck in doing the usual manual legwork in any step of the optimisation process. There are tools and methodologies available out there which can and will bring you to a near equilibrium to how you manage your campaign budget and resources.

Strategy & Decision Making
Set tangible goals and targets through software-aided rendering of all available data.

Basic Overview
Knowing what to do and when is a matter of thorough understanding of risks and opportunities coming from in-depth research. A typical SEO campaign may contain the following elements:

Phrase Research
Define best keywords by search volume utilising Google Keyword Tools, Webmaster Tools and Analytics. Refine results using sorting criteria designed to produce ROI-based phrase shortlist ready for campaign.

Competitive Analysis
This is the next step in the phrase refinement process which adds a layer of competitive strength to already defined phrases of interest in order to understand the likelihood and timeframes required to achieve rankings, traffic and ROI.

Targeting
By combining phrase research and competitive analysis we can produce a targeting matrix knowing exactly which set of phrases to match with what URL. This is essential for on-site optimisation activities and link building which are to follow.

Project Timeline
Defined timeline with expected activities, goals and outcomes to be noted each month of the campaign.

Senior Management: Expectations & Approach
Corporate SEO objectives can vary in nature so it’s important to understand how the management measures success. Setting their expectations right will avoid issues with budget and secure a healthy, long-term SEO
campaign. During your research you will discover many unique opportunities, however do not lose track of what matters to your superiors. They may have a completely different idea of a successful SEO campaign than you do and you do not wish to disappoint anyone.

You might find it hard to believe but others may not find your phrase research methodology that exciting. When you present your ideas and concepts use plenty of visual aids, comparison charts, summarise the message and always include tangible benefits.

Data Gathering: Look Everywhere & Record Everything

Arm yourself with as much data as you can as early as possible, you may need to chase different departments but stay persistent and remind everyone who owes you your data or access to relevant tools. This includes:

- Analytics Suites
- Google Webmaster Tools
- Raw Log Files
- PPC Campaigns
- Affiliate Campaigns
- Previous SEO Data / History
- Whitepapers & Case Studies
- Offline Content

Maintain good records and chronology of exported/accessed data. It’s also a good idea to write down departments, names and contact details of people responsible for data access. Take the initial snapshot of the entire state of the website and its performance in search engines. Archived data will ensure smooth comparative analysis at a later stage and may even help you present key benchmark when reporting to the management.
Your SEO budget is never unlimited, but others’ expectation of your performance may get pretty close to it. Arm yourself with some positive news right at the beginning. Small wins will help you get a bigger budget sooner.

Where to start?

If you work on a large corporate website, the odds are, it may already have a decent status in search engines and a smart thing to do would be to identify the easiest targets with the greatest success potential. Note that the ‘success’ can be measured in different ways (e.g. traffic, sales, leads, sign-ups) but none of that will happen unless the search engine position is where it should be.

Phrase Potential Calculation

At Dejan SEO we’ve utilise an in-house phrase targeting methodology based on potential calculation formula. We use up to two data sources (Google Webmaster Tools and Google Keyword Tool) which include search volumes, impressions, clicks, CTR values and average ranking positions.

Re-prioritised phrase targeting: Apply customised traffic and revenue projections to search volumes

The tool we use also enables us to add conversion rate and goal values to phrase potential calculation and produce estimated financial outcomes and projections based on both keyword-specific and site-average performance. This enables us to order our list of phrases and allocate our budget accordingly.

A wise thing to do is to allocate a fraction of your resources to what may be a set of easy targets while you
continue to chisel away at your long-term goals.

Technical Wizardry

Understanding on-site problems through near-magic.

Search engines cost a lot of money to run and they require incredible amounts of storage and processing power. The last thing they want to do is to go in navigational loops and index duplicate pages on a broken website. Yes, it’s often navigational and architecture problems that can act as bottlenecks in your SEO efforts. Google now employs various filters (e.g. Panda) to devalue websites with thin repetitive content.

Problem Visualisation

Most humans cannot easily comprehend plain data (CSV lists, XML sitemaps) but we can identify problems in seconds if visualised in the right way. Here’s an example of a website troubled by faceted navigation, starting from the top level sub-category:

In an isometric overview above we observe a single sub category breaking up into multiple navigational items, each splitting up into more up to 16 levels deep, each time generating more pages than before. If we attempted to show all categories at once we would be looking at a complex hypercube, not so useful. The problem grows even larger in a deeper linear analysis. Let’s zoom out a bit:
From the image above we can see that the navigational structure of this website leans towards infinity at an increasing rate. Although a healthy PageRank of 7 or 8 may disperse bots to, and index as many as 100 million pages this is not necessarily a good thing. This is not just because of search quality filters but also because of sub-optimal link equity distribution and intra-competitive pages within a single domain.

The visualisation above was generated using a desktop application called [3]PowerMapper, but the poor thing crashed before 'finishing' the infinite website structure. For more elaborate scanning you may wish to employ something more robust such as [4]80 Legs which utilises a network of machines and has an impressive set of customisable crawling parameters. The crawling data can be manipulated in any linked data processing application such as [5]Gephi and produce fine visualisations such as the one we did for one of our clients' websites:

The visualisation above illustrates the top linked pages within a single domain and its connections. The size and colour of each node is determined by the page in-degree or "internal PageRank". Using the same tool-set we can also project the internal promotion and importance of different pages and model outcomes of various
linking structures.

Note that there are many useful visualisation and mapping applications out there suited for different purposes so it’s a good idea to try out a few and discover which one may be most compatible with your data and requirements.

Visualisation is helpful indeed, but a holistic overview requires much more work than that. In the following diagram we see a typical breakdown of typical activities:

On-Site Review
Remember, things may move slowly and there could be long waiting periods until your on-site recommendations are implemented. There are often numerous items to fix from the technical point of view, but after forming a sound strategy you can prioritise your work and attend to high priority items first.

URL to Phrase Matching
This task is related to strategic targeting exercise and enables us to prevent any double-targeting or lack of URLs for any of strategically selected phrases.

Navigational & Canonical Setup
Ensures that there is always only one canonical version of each URL on your website and that navigation flows with optimum SEO efficiency.

Content Analysis
Checks for any content issues, accessibility of content and duplication.
Google Webmaster Tools Setup & Configuration
Main channel of communication between Google and your website. Enables us to easily detect any known issues and correct problems directly with Google.

Off-Site Review
Assessment of ranking factors not directly related with the technical setup or content located on the site itself.

Define Linkable Assets
Discover the most popular content on your website by observing links and traffic data.

Relationship Assessment
Learn about how you are connected in the business world and with your customers to extract value and potential link partnerships.

Broken Link Analysis
Finding out if any links are going to wrong pages or expired content.

Consolidation Potential
Discovering links which could be implemented better or online references and mentions that could turn into links.

Ongoing Activities
Looking after SEO performance on a month-to-month basis ensuring that your SEO runs smoothly, from technical issues and content to inbound links.

Implementation of SEO-friendly marketing policies
Much of your SEO can happen on the fly and without engagement of an SEO company. Try to point out areas where you can contribute to your SEO by adapting your daily marketing activities.

Link building campaign
Appropriate management of your link building campaign to secure a steady flow or links for your domain.

Content development
Development of creative content to serve as a marketing device, social booster and a link generator.

Risk assessment and monitoring
Understanding risks is essential to preservation and maintenance of your rankings for the long term. There are certain rules which search engines want webmasters to respect. Penalties can apply if those rules are broken.
4.4. April

Team Training

Start with a team-wide presentation introducing essential terms, concepts and activities. During this session all participants need to be familiarised with search marketing strategy, agendas and desired outcomes. The presentation should end with a Q & A segment. Each consecutive training session must be designed for each team or person depending on their role in the team (marketing, web design, programming, content, PR, advertising, social media, management).

During the training sessions each team member will learn on what they can do to help you reach you strategic search marketing goals.

Will this work? Honestly, no.

At first everyone will be eager and excited about SEO but you will soon learn that you can never get a complete cooperation from every department on a permanent basis. Do not get discouraged and use what you have available to you as there is no way around it.

2012 and Beyond

Here are a few emerging trends and technologies relevant to SEO:

Quality

Maintaining a lean, canonicalised and well laid out website is one piece of the puzzle. Google’s perception of ‘quality’ [8]goes deeper and they [9]attack manipulative techniques more aggressively each year. Ensure your rankings do not depend on purchased or rented links coming from questionable sources. The best way to do that is to perform an in-depth [10]link analysis and [11]look for unusual patterns.

Summary list of things to avoid:

- Index Bloat
- Thin, automated, repetitive content pages
• Indexable search results and tags
• ‘SEO’ keyword pages

Over-Optimise
• Anchor Text
• On-Page

Link Schemes
• Automated
• Blog Networks
• Spam
• Mass-Scale ‘Clever’ Tactics

Poor UX
• Navigation, Design & Layout

Google+
Sceptics argue that Google’s social platform will never take off, though there is [12]solid proof that [13]Google is more serious about their social venture than anything else before. They will do everything it takes to make it a success, and when they do go mainstream you will want to already have a well-established profile and broad presence with plenty of social endorsements (+1’s, shares, mentions, circles).

Structured Data
[14]Schema.org standard is not there just for shopping carts. There are many uses of structured data and search engines will [15]continue to advance and diversify the complexity of their results and aggregate data. For now, here’s one significant incentive to encourage you to jump on-board: click-through increase from rich-snippet based search results.

Authorship
[16]Authorship ties in with both search quality and Google+ and represents a significant element in both social and knowledge graph. In the future Google will continue to favour content from trusted authors and their social influence will be quite significant. Just like with schema.org mark-up, author snippet with a thumbnail increases the click-through rate significantly. Remember, just like with anything in enterprise SEO you need to ensure that authorship is embedded in the website’s CMS and guidelines created for authors and contributors on [17]how to verify their authorship on their end. You don’t want to do this manually each time.

Discussion
Tell us about your experience with enterprise SEO on [18]Google+

Downloads

Other Recommended Articles
If you like this article you might also want to read these:
4.5 May

- [21] Why Google Went Social

- [22] Search Quality: The Link Graph Theory

17. https://plus.google.com/authorship
18. https://plus.google.com/u/0/11407453274305880465/posts/drVogGReKzP

4.5 May

4.5.1 Google Knowledge Graph: How did this happen? (2012-05-17 14:16) - admin

Google goes one step further in attempt to become an intelligent search engine by introducing their knowledge graph results. "Search is a lot about discovery—the basic human need to learn and broaden your horizons.\textquotedbl", Google.
Predictions for 2012

Earlier this year I predicted the [1]three key improvements in Google’s algorithm will be improved spam
detection, social graph and knowledge graph. Two of my other articles outline ways of knowledge graph
SMX Sydney this year, I dropped in the concept of Google’s knowledge graph as the next big thing for

One thing most of us did not expect was for this to happen so soon!

Google’s First Announcement

Things, not strings!

One of the first public announcements on the next-generation Google search was made by Larry Page in the
”[6]2012 Update from the CEO” post on Google’s corporate blog:

"Understanding identity and relationships can also help us improve search. Today, most search results are
generic, so two strangers sitting next to each other in a café will get very similar answers. Yet everyone’s life
experiences are unique. We are all knowledgeable about different things; we have different interests and our
preferences—for music, food, vacations, sports, movies, TV shows, and especially people—vary enormously.”

"Google+ helps solve this problem for us because it enables Google to understand people and their connections. So when I search for Ben Smith, I get the real Ben Smith (for me), right there in my search box, complete with his picture. Previously, the search box would just have had the series of letters I had typed, with no real understanding that I was looking for a unique person. This is a huge and important change, and there’s a ton more work to do. But this kind of next-generation search in which Google understands real-world entities—things, not strings—will help improve our results in exciting new ways. It’s about building genuine knowledge into our search engine.”

We all knew Google was up to something since their acquisition of Freebase in 2010. Growth of their knowledge base from 12,000,000 to 200,000,000 nodes took place in a very short period of time hinting at the level of significance of the knowledge graph project to Google’s core product team. [8]Amit Singhal explains more on [9]Mashable and on his [10]Google+ post:

"In 2010, we acquired Freebase, an open-source knowledge graph, and in the time since we’ve grown it from 12 million interconnected entities and attributes to over 200 million. Our vision for this knowledge graph is as a tool to aid the creation of more knowledge – an endless cycle of creativity and insight.”

[11]Lance Ulanoff from Mashable has covered the latest update in a catchy headline calling Google "[12]1000 times smarter”.

Early Predictions
One guy who saw everything years ahead of everyone else was of course [13]Bill Slawski who first reviewed one of the related Google patents back in [14]August 2010. Today he points us at more related articles:

Google on Using a Knowledge Base of Articles to Make Searches Smarter

Google on Multi-Tiered Indexing and Multi-Staged Query Processing

Premature Feature? No way!

Since their recent [20]operational and product trim Google seems to be in an unstoppable roll-out of new features and improvements, more or less ad-hoc style, A/B testing things and trying stuff out. This shows that their new spirit of a startup company is not all about chasing Facebook’s tail but goes deeper into fundamental ways of how Google’s product development and deployment works. Good for you Larry Page.

First Glimpses

In early May, [21]Bas van den Beld from State of Search [22]started noticing the first signs of knowledge graph in action and later that day we collaboratively performed and published a [23]study on various query types which seems to be triggering Google’s answer style results above the usual ten organic results. Our favorite example is "[24]how old is Jesus".
Knowledge Graph: Right Now

Inside Search (Google) [26] talks more about the upcoming changes with Google’s search results and gives some nice examples of where these can be seen as useful for users:

To find out more how and when the new search results may appear we recommend [27]this article by Google.

Google Research Tool

[28] Google research tool has [29] quietly entered Google Docs and now gives users an option of researching facts while writing. When writing a new document Google will now display the research pane which contains a search bar and series of organic results. One handy feature we spotted was the ability to quickly link to quickly preview relevant results, link to them or include properly formatted citations within your article. Quite a time saver.
For now, the functionality of the research tool is limited to google.com organic results including image results, but we anticipate to see Google’s knowledge graph data in there soon as well as support for more localities.

Google Knowledge Graph Video

References:

- [33]http://mashable.com/2012/02/13/google-knowledge-graph-change-search/
• [40]https://support.google.com/docs/bin/answer.py?hl=en &answer=2481802
• [41]http://support.google.com/websearch/bin/answer.py?hl=en &answer=2620861

10. https://plus.google.com/u/0/11574399689614835150/posts/3vLRVL7C4QS
12. http://mashable.com/2012/05/16/google-knowledge-graph/
27. http://support.google.com/websearch/bin/answer.py?hl=en&answer=2620861
28. https://support.google.com/docs/bin/answer.py?hl=en&answer=2481802
30. http://www.youtube.com/embed/mmQl6VGvX-c?rel=0
33. http://mashable.com/2012/02/13/google-knowledge-graph-change-search/
4.5.2 PageRank Modelling Study (2012-05-22 11:50) - admin

This is an academic study of controlled PageRank modelling of a simplified link graph in a collection of three documents.

Key Observations:

1. Introduction of a new document into the existing collection reduces the PageRank value of existing documents.
2. Differences in PageRank value of chain-linked documents vs. directly linked documents.

Experiment Parameters:

- Graph Type: Directed
- Epsilon: 0.2 (default: 0.001)
- Probability: 0.85
- Use edge weight: Yes

Definitions:

- 'Node' = page.
- 'Edge' = link, can have two directions and three states (a->b, b->a and a<->b)
- 'Link Graph' = collection of nodes and edges

The Test

Using the traditional formula we calculate the distribution of PageRank between two nodes connected with an a->b edge resulting in the initial PageRank split of:

- Node a: 0.33
- Node b: 0.66.
Introduction of a completely new node in existing the collection (in our case the three nodes and one directed edge represent the entire web). Observe the change in PageRank allocation.

- Node a: 0.27 (old value 0.33)
- Node b: 0.47 (old value: 0.66)
- Node c: 0.27 (new node)

An edge between Node c and Node a will pass fraction of the Node c PageRank which will in turn pass the
extra value to Node b.

The outcome:

- Node a: 0.35 (old 0.27)
- Node b: 0.45 (old 0.47)
- Node c: 0.2 (old 0.27)
PageRank modelling for the purposes of this study was calculated using Gephi.

4.5.3 Survey Results: Users Prefer Post-Penguin Google (2012-05-29 21:01) - admin

We asked 2,190 Google users if they noticed any changes in Google results since Penguin. This is what they said.

Google implements countless changes and tweaks to their search results over a single year and most users don’t notice more subtle changes such as order of websites in search results. March and April, however, brought series of algorithmic improvements which were even announced on Google’s blog hinting at their impact and significance. On the other hand SEO community has been quite stirred up and even divided with recent algorithm updates targeted at search quality and web spam.

Out of 9,610 survey views we received 2,190 responses within 24 hour period. Data was collected from news (47.4 %), reference (34.1 %) and arts & entertainment (18.4 %) publisher categories. Among people earning $25-49K, those aged 25-34 picked Yes 2x more than those aged 55-64. We used Google Consumer Insights to ask Google users (USA only) if they recently noticed any changes in the quality of Google’s results.

Was Penguin good for the user?

Those who noticed changes in Google results were asked to describe whether in their opinion Google’s search quality has increased or decreased and by how much. We offered a 1-7 star rating selection purposely in order to capture fake positives in the middle (4). In total we collected 500 responses and here is the breakdown of answers.

How would you describe Google results now in comparison to 30 days ago or before?
Note: 24.4% users answered that Google’s search is neither better nor worse, in that case they should have answered the qualifying question with “NO” and end the survey there.

Factoring in all 500 answers the average response tends to lean towards neutrality (with an average of 4.4) with a slight hint of possible improvement in users perception of Google’s search result quality. Distributing the value of (4) evenly across both ends (positive and negative) we end up with 41% of those not satisfied with recent changes and 59% who feel the results are now better.

Conclusion
Most Google users didn’t notice the change at all, however those who have, feel that Google now shows higher quality results than before.

Survey Data:

- [5]Results by Gender
- [6]Results by Age
- [7]Results by Income

4.6 June

4.6.1 Correlation Equals Causation (2012-06-14 01:11) - admin

I have to say that I am a bit fed up with all this "correlation does not equal causation" talk and wish to bring something new to your attention.

Authors of awesome articles rich in research, analysis and data, often feel compelled to inject the correlation ≠ causation clause. It should be obvious that we are all students of Google’s search and not its engineers and I would imagine it’s implied that we’re all making our best guess at what goes on within the algorithm. Are there really any SEO professionals out there who actually believe that one site ranks higher than the other purely because its mozRank is higher by 2 points? Perhaps, but that’s a totally different story.

[blockquote type="blockquote quotes" align="right"]We take the top 100 sites ranked by our algorithm and compare it with the rank on technorati.com. Out of our top 100 blogs, 49 were listed by technorati.com. Of these 49 blogs, 40 % had ranks less than 50 in technorati.com, 53 % had ranks less than 100, and 71 % had ranks less than 200. This suggests a very good correlation. Table 2 shows the top 25 urls ranked by our algorithm. The “NA” is used in cases in which technorati.com did not provide a ranking.[/blockquote]
Correlation = Causation

Here’s something people don’t consider at first, and that is that Google relies on correlation principles and considers correlation in research and within their ranking algorithm. Their observations apply across many different facets of search including: content quality, user behaviour and page characteristics. Existing algorithms are often validated through sandboxing and limited public releases where correlation is observed prior to global algorithm update.

In a [3] paper published in 2011, Google researchers observe behaviour of viral videos and attempt to design a ranking formula for viral video blogs. As part of the ranking function validation they look at websites such as technorati.com, www.huffingtonpost.com and www.gizmodo.com.

Correlation has also been considered with Google’s study on [4] impact of ranking of organic search results on the incrementality of search ads as well as in their [5] analysis of attractiveness of search results through presentation variations such as title term bolding.

Correlation, Panda & Page Layout Algorithm

One of the more obvious examples are series of algorithmic improvements Google rolled out in 2011 and 2012 which were primarily aimed at page quality. So rather than being designed in the lab and released in the wild, the algorithm was designed with help of its representative users and their feedback.

For those of you not familiar, here are some of the questions quality raters / control group could have been asked (from an article by [6] Amit Singhal):

- Would you trust the information presented in this article?
- Does this article have spelling, stylistic, or factual errors?
- How much quality control is done on content?
- Does the article describe both sides of a story?
- Is the site a recognized authority on its topic?
- More ...

I would say that a site which answers favourably all of the above would likely rank high in Google. Naturally they didn’t get it right from the start and there were many innocent victims of various algorithm updates, yet the process continues to be refined and correlation to result quality observed.

Conclusion

My objective here is not to dispute basic scientific principles. Instead, I wish point out that perhaps we’re starting to create a vibe of negativity and lack of trust in our own findings and data in fear of being incorrect.

Fear not and embrace your correlation and apply generously - but do apply common sense.

4.7 July

4.7.1 SEO Experiment With Non-Link References (2012-07-11 23:36) - admin

Does Google use written (non-linked) URLs as a ranking signal? Dejan SEO team investigates with a simple experiment.

The Almost Link

The web is full of plain text URL references and to me they are the missing link in the link graph.

Consider the differences between text links and text with written URL. One contains an `<a href="` and the other does not. What about the intention of the author? One would think that it's nearly the same. Why would anyone write down a URL if not to instruct their readers to follow it?

Where do we see plain URLs?

- Academic references
- Document footnotes
- Word documents
Authors sometimes do not link even in HTML documents where it’s easy to do so. Whether it’s a matter of being lazy, rushed or not having the knowledge or CMS to support it is irrelevant.

Examples:

- for more information on hedgehogs visit: http://en.wikipedia.org/wiki/Hedgehog
- a Wikipedia article on hedgehogs (http://en.wikipedia.org/wiki/Hedgehog) might be a good place to start your research
- http://en.wikipedia.org/wiki/Hedgehog (all you need to know about hedgehogs)

So our idea is that Google surely uses plain URL references as a signal comparable to a hyperlink with anchor text being subsidised by surrounding text.

Test: Stage 1

We registered a brand new domain (never registered before, clean history) and put up a single page of content with a phrase we wanted to rank for.

One articles was posted mentioning the new domain URL but not linking to it. Phrase of interest was placed in various places including document title, image tags and in the proximity of the ‘target’ URL itself.

Outcome: Google found and indexed the domain. No movement in rankings.

Test: Stage 2

Each day we placed a new article on a different domain and monitored the outcome in the rankings. After ten new references we stopped and waited for a week.

Outcome: No movement in rankings.

Test: Stage 3

Personal note: I was really disappointed at this point as I expected these references to do at least something visible, but nothing happened.

We repeated the process with additional ten domains and articles.

Outcome: All 20 articles were indexed and cached. No movement in rankings of our new domain.
Conclusion

There is no solid evidence that written URLs act as anything more but as an equivalent of a written word on the page helping the page itself rank but not the URL it refers to. We have not tested the topical influence of domains already ranking due to existing inbound links.

One thing is for sure though, Google does discover new URLs and visits them, even if they’re simply written down and not hyperlinked.

Could we be wrong?

There is one (unlikely) possibility that the test domain may have moved up, but was consequently moved down by the referring pages which were added to Google’s index daily, resulting in 20 new results pushing the test domain down in rankings.

What Matt Cutts said in 2009:


Response to Reader Feedback

Michael Martinez, adds some great thoughts on how our experiment could have gone wrong and talks about challenges of running SEO experiments in general. For those not familiar, Michael has already spent a great deal of time considering various challenges related to attempts to reverse-engineering Google’s algorithm.

My reply and further thoughts to his feedback are below.

Your test is flawed for several reasons, not the least being that you were trying to rank a newly registered domain (which has no history in Google’s index) with your test signals. Yes, new domains are registered all the time – but you’re only looking at one part of the picture. Signals of all types can have a greater impact for well-established Websites.

No doubt, there may have been a different impact on an established website but due to same signal noise reasons we chose to go with a brand new domain.

We had one unsuccessful test of this kind in the past. At the end of the test we directed a single link to our target domain. It shot up to the top as soon as the page with the link was indexed. Needless to say we picked a really easy, long-tail phrase.

It’s almost completely mathematically impossible to reverse-engineer any specific Google ranking signal anyway. The algorithm doesn’t allow you to isolate any specific factors so how can you identify them?

The only thing we can do is minimise the signal noise and run clean simple experiments where a number of variables is reduced to a minimum.

The best you can do is to look at a NATURAL search result (an existing, real query that is already fully populated by natural, unoptimized content) and then ”shake up the mix” with a simple change to see what...

abiclone.com.au/blog
happens.

This was discussed internally on many occasions but we lack both time and funds to run an SEO test as elaborate as that. These types of tests appear simple at first but as you continue to plan realise just how many properties everything has.

But even behind a natural query there are algorithmic factors that may prevent a lot of your tests from having any impact, even though they might have an impact in highly competitive queries.

Care to share some examples of the type of algorithmic factor which may prevent tests of this type?

As for reporting, you don’t provide any detail about the age or quality of the Websites where you placed the articles, nor about when the articles were cached in Google, nor whether the articles were passing control anchor text through normal links to other destinations, etc.

- Domains are 5 years old in average (range from 2 to 8).
- Articles were indexed and cached in Google
- They also ranked in Google for the term we wanted the target domain to rank for
- There was no control anchor text through normal links to other destinations (great idea!)

Hence, it is impossible for anyone to replicate your experiment and compare their results to your reported results. At best you have provided one more anecdote for people to think about but your experiment is not very useful beyond stimulating thought and conversation.

I would love to see a follow-up experiment so if you need any more data from me please let me know. I will share if possible.

Hypothetically Google could be following these URL references only for "discovery", meaning they just want to include the destinations in the index. Of course, Google often indexes new domains without any links or references to them at all.

Yes. Google discovers domains in numerous ways, but I did get a confirmation from one of Googlers on this particular situation. They will discover a page by following a URL reference, but will not count it as a link signal.

In my opinion there is no doubt whether they use it as a signal in one way or the other, but like you and a few others said, it could be either a subtle addition or a blend of various complementing signals which work in 'teams'.

One additional hunch I have about our inability to reverse-engineer Google’s search results is the fact that Google may be purposely randomising results to a degree. This would not only serve as a protection mechanism but as a useful device to keep the results fresher and new content discoverable. I see it as kind of
a DNA mutation as part of search evolution where natural selection is subsidised by user choice.

1. http://www.youtube.com/embed/nlByfISl35w?rel=0

4.7.2 Detecting Google Search Layers with Advanced Search Operators (2012-07-15 14:04) - admin

The way Google presents search results has always been and always will be a bit of an enigma to all of us. Every now and then, however, a snippet of information slips through the cracks giving us a chance to take a look at its search anatomy. In the example below, we can see a search for "greater than 100 years old", or "[1]100.. years old". There are four exceptions in this particular example where the operator did not work as expected:

1. Wikipedia

2. News Results

3. PDF Documents

4. Movie Database

News Results were not a surprise, but the other three are. This perhaps tells us that Google treats, stores and processes different types of resources in different ways.
This is just a preliminary test and we’ll continue our investigation to get more clarity, especially around Wikipedia entries.

1. [https://www.google.com.au/search?q=100+years+old&pws=0](https://www.google.com.au/search?q=100+years+old&pws=0)

### 4.7.3 How I Accidentally Conducted My First Panda Experiment (2012-07-26 00:47) - admin

Taking a look at an old search experiment with some interesting observations in the light of recent algorithm changes.
Note: I welcome all input, feedback and constructive criticism. However, if you’re of the opinion that SEO experiments are pointless, then please leave now before you get deeply upset by the ‘speculative’ nature of this post.

Test #1

In early February 2011, two domains were registered:


Both were linked from the [3]same page in exactly the same order as above. For results and the technical set-up of the experiment you can [4]read more here.


Why did hyphenated domain outrank the non-hyphenated domain?

The main differences between the two sites:

1. Hyphenated domain was linked first on the source page

2. Hyphenated domain was indexed first

3. Hyphenation (claimed to be a negative factor)

Clearly, the number of hyphens in the domain did not prevent it to outrank the second domain in our experiment.
Test #2

At around the same time I noticed a pattern of pages displayed in search results for the site: command and wanted to see what’s going on in a controlled environment. So we used the same domain [6]http://dejanseolovestesting.com and set up a few pages on it. Each page was of a different URL length and [7]they were all linked from the same source and in the correct order.


Results

Date: 10th August 2011
As visible from the search query ([site:dejanseolovestesting.com]) the results are sorted in the following order:

1. INDEX
2. 1234567
3. 123
4. 12
5. 12345678
6. 12345
7. 123456
8. 123456789
9. 1234
10. 1

Expected outcome against actual order:

- 7:1234567
- 3:123
- 2:12
- 8:12345678
- 5:12345
- 6:123456
- 9:123456789
- 4:1234
- 1:1

Observations: There seems to be an underlying pattern with numbers funneling down and re-setting at certain points.

Pattern: 7,3,2,8,5,6,9,4,1

We followed up to this search query and found the pattern unchanged. Nobody has yet been able to figure out why this particular order. Perhaps an independent experiment may give an answer whether this same pattern would repeat or not.

So what’s Panda got to do with this?
Well, Panda 1.0 was released in late February 2011, just after my tests had finished. One domain has a single page in index, the other had a whole bunch of pages - all with very thin content.

I am starting to think that I may have actually tested Panda filter in action for the first time back in February last year. Without realising what I was doing.

Thoughts?

10. https://plus.google.com/114074532743058808065/posts/2np6B3wZDQ1

4.8 August

4.8.1 Unusually Fresh PageRank Update in August 2012 (2012-08-05 02:25) - admin

Google’s toolbar PageRank (TBPR) values were released to public in August this year, but with much fresher snapshot date than we remember.

PageRank value is always a historic figure showing real-time PageRank score of a page at the time when the snapshot was taken. There is also a delay between the day when snapshot is taken and when TBPR is pushed out to public. The graphic below illustrates changes in time between snapshot and push, and between public updates. Right now PageRank is just under two weeks old but it will become more stale and inaccurate as we move closer to the next public update. We can say that PageRank is most inaccurate just before each public update.

After going through all available data I had on PageRank updates I made this graphic which illustrates changes in PageRank freshness.

It certainly seems more regular and quite stable:
What we see, the minimal lag between snapshot and public release is not a new thing and there is a clear pattern of fresh PageRank persisting since July 2011.

How do I determine the lag value?
Each PageRank update I go and look for the earliest post on my blog I can find that is now showing PageRank. I use that as a rough benchmark to gauge when the PageRank snapshot was taken. Many freshly published pages show PR zero even after the update indicating that they did not exist at the time the snapshot took place.


4.8.2 The Random Google (2012-08-18 18:15) - admin

Is Google purposely randomizing things just to mess with us? Let’s take a look at one patent and two SEO experiments.
In March 2011, I wrote an article outlining an interesting phenomenon I observed with PageRank distribution in a controlled test environment. Linked to a number of similar instances it made me believe that there is an element of randomness in much of what Google does in search.

"The apparently random element in content indexation and assigning of PageRank which in addition to variable value of PageRank causes indexation to suddenly stop and resumes at later stage (perhaps when crawlers internal limit has been reached) appears to be a complex set of variety of internal rules rather than a purely random behaviour.

Some degree of random behaviour could potentially be beneficial to a search engine as it would provide a platform for organic growth of the algorithm much like sporadic mutations in gene replication enable living species to evolve.

Secondary benefit could be a layer of protection against deliberate reverse-engineering of its algorithm as each attempt to probe the system would result in subtly distorted version that would not match previous attempts."

Cited from: [1]Obstacles in Experimental Testing and Reverse Engineering of Google Algorithm

One month later we did a [2]domain hyphenation test which apart from its own results revealed another element of randomisation in Google’s listings.

Here’s a screenshot:
We of course expected these pages to list in proper order.

The second experiment was conducted a year after, this time with images and the results were even more baffling:
What should have been an ordered list of images appearing in a neatly organised list on the source page is randomized in the search results. Order was dictated by file name, image content, alt and position in HTML. My comments after observing this were again pointing at a possibility that random has a role at Google:

"Google may be purposely adding a RND function to things to mix things up a bit. This would keep results fresher and allow random discovery. Also it would prevent any systematic attempts at reverse-engineering of Google's ranking algorithm."


We performed a similar, [4]simpler test with two variations, and the results were also [5]mixed up and out of order in [6]both cases. Order on the page, nofollow or other devices we attempted did not help with getting the results display what we wanted.

That was supposed to say "SEO IS FUN".

Finally Bill Slawski [7]spots a [8]patent which describes algorithmic behavior designed specifically to confuse and confirm those attempting to game with Google's algorithm:

"A system determines a first rank associated with a document and determines a second rank associated with the document, where the second rank is different from the first rank. The system also changes, during a transition period that occurs during a transition from the first rank to the second rank, a transition rank
associated with the document based on a rank transition function that varies the transition rank over time without any change in ranking factors associated with the document.”


Have you observed any cases of randomization in Google results?

References:


4.8.3 Small-Scale CTR Impact on Search Results (2012-08-23 13:01) - admin

We tested the influence of small-scale click-through-rate on position of two pages on two new domains in Google. Two freshly registered domains (A & B) were used.

What did our test involve?
1. Search Google for "keyword"
2. Click on domain B in results.

Test Length: 2 Weeks
Participants: 100

Both domains were manually checked on a daily basis for position in Chrome incognito mode by the same user and location.

Here’s what we saw:
Alpha domain was indexed faster and they generally moved up and down in tandem, until day 10 which is where Beta domain took lead (with one sudden drop). The main movements are likely to major forces at play where as more subtle ranking changes could be due to CTR difference. At this point we conclude that small-scale CTR difference makes no significant impact, and the difference in rankings is likely due to a random link or reference appearing elsewhere.

We intend to repeat this experiment with more participants and longer period of time, if anyone else wants to join in or run one in parallel please let us know.

Position Table:
Alpha Beta 72 300 64 65 70 71 73 171 179 69 71 66 71 64 69 116 115 115 76 121 79 113 287 124 79

4.9  September

4.9.1  Google Plus SEO Audit (2012-09-10 17:30) - admin

Learn from the best they say, so I decided to run an SEO audit and analyse the architecture of Google+ to see if I can learn anything useful.
Canonicalisation

Google+ has a number of canonical issues and I will highlight one of them. On my posts page there are 11 instances of links to /u/0/

- [1]https://plus.google.com/u/0/

This causes everything to have two URLs with parameters which don’t change the content of the page. Something webmasters are advised to control through rel="canonical" and/or parameter settings in Google Webmaster Tools.

Here’s an Example:


What’s amusing though is that anything that links to a noindexed directory goes in, even uppercase/lowercase variations as illustrated above.

Bonus: You can also change the 0 to 1 in the URL and [7]double the canonical trouble visible in [8]this search query:


Coca-Cola:

1. https://plus.google.com/+Coca-Cola


UPDATE: Source code on /posts page contains <link rel="canonical" href="[14]https://plus.google.com/+Coca-Cola"> and canonicalises to relevant pages on /about, /photos and /videos.

Unrelated to this audit, but still amusing, is that when you search for Coca-Cola you find Pepsi (who don’t even have their vanity URL yet):
What is interesting is that the "Co" in Pepsi’s page has been triggered as a relevant search term to "Coca Cola". Why?

Reader Comments and Suggestions

[15]

[16]W.E. Jonk
On canonicalisation you can also add the mobile site which has a disallow: [17]m.google.com/app/plus

@[18]dejanseo As far as i’m aware /u/0 and /u/1 are used to control which account you want to view if you’re logged into multiple G accounts.

— Mark Hannant (@tripleox) [19]September 10, 2012


Well-spotted.

Duplicate Content & Hidden Links?

If you investigate links to my Google+ profile you will see the [21]following list:
Case 1: I +1’d Dan Russell’s post and my link is nofollow, I am not sure why this shows up in the link: if it’s not part of Google’s link graph. In any case I accept that there is a connection.

Case 2: The hangout post links to my profile directly as I was one of the participants. No problems there.

Case 3: This one was confusing as it had nothing to do with me. It’s a post about a car to which I did not engage in any way. What’s going on?

On closer inspection I found that the source code of the cached version of Lars Fosdal’s post also contains content from his recent posts and shares including one of my posts. What’s interesting is that the text-only version does not! Now all this wouldn’t be that unusual if it wasn’t for the fact that this content is not actually meant to be loaded and displayed. The link pointing to my profile is basically hidden in javascript googlebot has figured out not only how to get to the URL but the URL is also visible in the link: command.

Here’s how it looks like:

”[“Dan Petrovic”,”111588754935244257268”,1,1,”https://lh5.googleusercontent.com/-c-qbWzoKUA0/AAAAAAAI/AAAAAA/A/MjzYOJzpxal/photo.jpg”,”.1/111588754935244257268”]"
So how’s this duplicate content? Well, if the same snippet is contained on a number of recent posts in addition to the posts page and the post URL itself we’re looking at more than 3, possibly 10-20 duplicated chunks of content. Now, clearly the 'hidden links' are not there for malicious or manipulative reasons so I’m letting Google of the hook for that one. At this stage I am more concerned about their content duplication issues.

Sitemaps

Google+ sitemaps have been broken up into small chunks ranging from 0 to 47,999 and referenced from the [27]robots.txt as http://www.gstatic.com/s2/sitemaps/profiles-sitemap.xml


I am not sure about their scheduling but given the amount of new users joining their social platform one would expect fresher data:

<lastmod>2012-04-20</lastmod>

As you can see their last modified date is set to 20th of April 2012.

Another unusual thing is their referencing of URL pages:
It could be that users with "profiles." have not actually joined Google+ yet. I attempted to access the .profiles page with a plus. URL and noticed a neat 301 redirect back to the .profiles URL.

PageRank Sculpting

Google’s advice to webmasters is not to worry about PageRank sculpting and that using rel="nofollow" on internal links is time wasted on something which could have produced better results, for example write good content or engage on social networks. What puzzles me is why Google+ then uses nofollow or many of their internal links and what exactly they are trying to achieve with that.

This is my question for Matt Cutts, if you’d like it answered please vote for it on Google Moderator:

Google+ is using rel="nofollow" on some of its internal links. Why?

"Sometimes we try to be cautious, just so we don’t become any sort of spam attack vector ourselves. Over time, as we come up with ways to make sure that things are spam-resistant, we might remove that nofollow attribute though." Matt Cutts, Mountain View

Although I do appreciate the answer I still don’t understand how this would help since they’re not using nofollow on external links and using it on only selected internal links. It just doesn’t strike me as an anti-spam prevention measure.

I started mapping the architecture of Google+ in order to understand why they did this and have initially come up with a few modes of various entities and actions on their social network:

Activity
Item
Ownership
Plus Image Own Share Video 3rd Party Mention Post Hashtag Comment Tag Event Comment Page Add Userpage Invite Circle Review URL Place Hangout

In the next stage I started mapping various scenarios starting with posts, both original and shared.

Posts
When you create your own post links will flow naturally unobstructed by nofollow, however when people comment, +1 or share your post, link to their profile will be nofollowed. This applies to both users and pages. Exception to the nofollow rule are +Mentions which flow naturally as well as any links shared in post comments, including links to external sites which is odd. It seems Google is not afraid of linking out and counting on the post author to moderate.

Shared Posts
Shared posts operate similar to original posts except that the original author’s profile link gets a nofollow. Why? To be fair the link to the original post is a dofollow.
Photos

Photos and images have two distinct modes. They can appear in context of a post (either as tagged, original and shared) and as a gallery item. Tagged people are nofollowed in the post mode, but like most other things are dofollow in the gallery mode. In gallery mode live links to external links are not possible. Pasting a URL in the gallery-mode comments will simply display it as text and not include the hyperlink.

Examples:

- Image Share Post: [31]https://plus.google.com/111588754935244257268/posts/JpSWtTDbiNT

Videos

Essentially the same thing as photos, in both shared post and gallery mode. Shared author gets a nofollow, original post is followed, hashtags, +Mentions and live links are also followed.

In the [32]gallery mode for this video URLs are not hyperlinked and everything that is linked up is dofollow.

Comments

As you’ve already seen from examples above, comment authors are followed links only in media pages (images, videos) in gallery mode. Everywhere else they’re nofollowed. Comment +Mentions and URL do not contain nofollow.

Events

Events are nofollow free, posts, comments, mentions - the works. Here’s an [33]example.

One thing that confuses me about event posts is the two similar post types created with the event. One is supposed to be an actual post (it contains ripples option, event description and the ”on air” button) and the other one is the event itself in form of a post where people can interact (RSVP) their attendance and invite others.
Followed links on event posts (apart from the usual ones) are selected three who have set their attendance to "Going". Neither of them have a post URL you can easily grab as you would with everything else by clicking on the post date. Instead the URL is available once the event post is re-shared. How strange!

Here you can see examples of shared event and event post which now have available URL:

- Shared Event: [34]https://plus.google.com/111588754935244257268/posts/cxgYrAHadMD (only visible to my extended circles, you’ll have to [35]add me to see it)

- Shared Event Post: [36]https://plus.google.com/111588754935244257268/posts/My9vQhr7eA (available publicly)

Note that one of them doesn’t have the original "post" link reference whereas the other links to the [37]event page instead.

Hangouts
Hangout participants listed on a public post are all followed profile links:

Circles

Links to people in shared circles are nofollow:

Pages & User Profiles

There are three key areas I observed with pages and user profiles on Google+. First one is the fact that the tabs: posts, about, photos, videos...etc basically all represent [38]one page cached in Google and almost simulate URLs for the purpose of usability.
Basically, what I’m saying is that at least two of the above URLs in orange are essentially the same page in Google’s cache (posts and about). Further to that the followers and following segments (in red) doesn’t seem to be in cache nor source code so this means that adding people for the purpose of manipulating PageRank will not work. Finally the page’s official URL (in green) is an enabled link which makes sense.

Reader Observations

[39] Lyndon NA - It’s interesting that G handles the URLs as "real", yet appears to utilise JS to render the relevant part. Technically it would seem that you have URLs with dupe mark-up/content, and a bit of cloaking.

Place Page

Engagement on the place page is rewarded by a dofollow link for reviewers who have a Google+ profile activated. The rest are [41]not hyperlinked.
This article may be updated again soon.

[42] Drop me a line if you find something interesting and I’ll add you as the article contributor.

1. https://plus.google.com/u/0/
2. https://plus.google.com/
3. https://plus.google.com/+Coca-Cola
5. https://plus.google.com/s/SEO
By factoring in bounce rate data in calculation of PageRank values, Fabien Mathieu and Mohamed Bouklit create a new and enhanced ranking algorithm.
In their research paper titled "The Effect of the Back Button in a Random Walk: Application for PageRank", Mathieu and Bouklit (LIRMM) propose an enhancement to Google’s original algorithm. They factor in the bounce rate of a document and refer to it as either "Reversible" or "Irreversible Back". By doing so they’re giving the traditional PageRank model an interesting and potentially useful addition, a possibility for the random surfer to "return" by hitting the back button.

This is a fundamental shift in the well-established PageRank model which has found its application in numerous areas already, even beyond search. The concept of "backoff process" was first introduced by Fagin in "Random walks with back buttons" in 2000.

Mathieu and Bouklit take this a step further and apply their formulas to a collection containing 8 million documents which has produced results different to those of typical Google results. The research paper was published in 2004 and it is likely that their version of PageRank which offers better modelisation of the web users has been merged with a semantic pertinence-sort and tested out in a practical search situation.

We will attempt to contact the authors of this paper for their comments.

References:

4.9.3 Authorship: Messing Up Your Rankings (2012-09-30 13:15) - admin

Your ranking data could be skewed since the introduction of authorship suggestions. In this article we investigate how this could happen.
Major SEO portals have recently covered the "hidden benefit of authorship" where users who return back to results after reading your article are presented with additional three articles written by you. We investigated this feature to learn more how Google selects what to show, and the answer is simple: they use topical relevance as the first and major factor (as opposed to +’s, shares, PageRank and date which could play secondary role in the selection).

Now, the interesting part is that if you break down the URL of any of the suggested article links you will see the following:

1. url?sa=t
2. &rct=j
3. &q=
4. &esrc=s
5. &source=web
6. &cd=1
7. &cad=rja
8. &ved=<hash>
9. &url=<URL>
10. &ei=<hash>
11. &usg=<hash>

Notice the CD=1? This is the part which could be causing trouble in your ranking data within Google Webmaster Tools.
If users click on the first of the suggested articles it will be treated as a first position, but for what phrase? If it’s the original phrase then this might not be the problem as Google Webmaster Tools show the position for the highest ranking page. Consider the “Top Pages” data however, and how this may impact the average position for specific URLs. It could for example make them appear to be ranking higher than they are if Google users click on the first suggested result which is normally found on position 9 or 10, or even on the second page of results for most searches. Similarly a top ranking page could appear to be ranking lower if it ends up being a third suggested result to your other content, passing on the CD=3 value.

At this stage we’re unsure whether Google uses the CD= variable to feed the data into Google Webmaster
Tools too or through a different mechanism, but this is definitely one thing that’s worth keeping an eye on. We’ll be asking this question in the next hangout with Google in hope to clarify this issue.

4.10 November

4.10.1 How I Hijacked Rand Fishkin’s Blog (2012-11-06 18:09) - admin

Search result hijacking is a surprisingly straightforward process. This post will go over theory, test cases done by Dejan SEO team and offer ways for webmasters to defend against search result theft.

I wish to thank Jim Munro, [1]Rob Maas and Rand Fishkin for allowing me to run my experiment on their pages.


Brief Introduction

Before I go any further I’d like to make it clear that this is not a bug, hack or an exploit - it’s a feature. Google’s algorithm prevents duplicate content displaying in search results and everything is fine until you find yourself on the wrong end of the duplication scale. From time to time a larger, more authoritative site will overtake smaller websites’ position in the rankings for their own content. Read on to find out how exactly this happens.
Search Theory

When there are two identical documents on the web, Google will pick the one with higher PageRank and use it in results. It will also forward any links from any perceived 'duplicate' towards the selected 'main' document. This idea first came to my mind while reading a paper called "Large-scale Incremental Processing Using Distributed Transactions and Notifications" by Daniel Peng and Frank Dabek from Google.

Here is the key part:

"Consider the task of building an index of the web that can be used to answer search queries. The indexing system starts by crawling every page on the web and processing them while maintaining a set of invariants on the index. For example, if the same content is crawled under multiple URLs, only the URL with the highest PageRank appears in the index. Each link is also inverted so that the anchor text from each outgoing link is attached to the page the link points to. Link inversion must work across duplicates: links to a duplicate of a page should be forwarded to the highest PageRank duplicate if necessary."

Case Studies

I decided to test the above theory on real pages from Google’s index. The following pages were our selected ‘victims’.

1. MarketBizz
2. Dumb SEO Questions
3. ShopSafe
4. Rand Fishkin’s Blog

Case Study #1: MarketBizz
26 October 2012: Rob Maas kindly volunteered for the first stage test and offered one of his English language pages for our first ‘hijack’ attempt. We set up a subdomain called rob.dejanseo.com.au and created a single page [4]http://rob.dejanseo.com.au/ReferentieEN.htm by copying the original HTML and images. The newly created page was +’ed and linked to from our blog. At this stage it was uncertain how similar (or identical) the two documents had to be for our test to work.

30 October 2012: Search result successfully hijacked. Not only did our new subdomain replace Rob’s page in results but the [5]info: command was now showing the new page even for the original page and it’s original PageRank 1 was replaced by PageRank "0" of the new page. Note: Do not confuse the toolbar PageRank of zero with real-time PageRank which was calculated to be 4.

Notice how the info: search for the URL returns our test domain instead?
4.10. November

So all it took was higher PageRank stream to the new page and a few days to allow for indexing of the new page.

Search for text from the original page also returned the new document:

One interesting fact is that site:www.marketbizz.nl still returns the original page "www.marketbizz.nl/en/ReferentieEN.htm" and does not omit it from site search results. Interestingly that URL does not return any results for cache, just like the copy we created. Google’s merge seems pretty thorough and complete in this case.

Case Study #2: dumbseoquestions.com

30 October 2012: Jim Munro volunteers his website dumbseoquestions.com in order to test whether authorship helps against result hijacking attempts. We copied his content and replicated it on
1 November 2012: The next day Jim’s page was replaced with our subdomain, rendering Jim’s original as a duplicate in Google’s index. This suggests that authorship did very little or nothing to stop this from happening.

The original website was replaced for both [7]info: command and search queries.

Interesting Discovery

Search for the exact match brand ”[8]Dumb SEO Questions” brings the correct result and not the newly created subdomain. This potentially reveals domain/query match layer of Google’s algorithm in action.
Whether Jim’s authorship helped in this instance is uncertain, but we did discover two conflicting search queries:

1. [9]Today we were fortunate to be joined by Richard Hearne from Red Cardinal Ltd. (returns the original site)

2. [10]Dumb+SEO+questions+answered+by+some+of+the+world’s+leading+SEO+practitioners (returns a copy)

One returned the original site while the other showed its copy. At this stage we have not yet tested the impact of rel=”canonical” in potential prevention of result hijacking and for that reason we created a separate experiment.

Case Study #3: Shop Safe
The following subdomain was created [11]http://shopsafe.dejanseo.com.au/ replicating a page which contained rel=”canonical”. Naturally the tag was stripped off on the duplicate page for the purposes of the experiment.

This page managed to overtake the original in search, but never replaced it when tested using the [12]info: command. All +1’s were purposely removed after the hijack to see if the page would be restored. Several days later the original page overtook the copy, however it is unclear if +’s had any impact on this.

Possible defense mechanisms:

1. Presence of rel=”canonical” on the original page

2. Authorship markup / link from Google+ profile

3. +1’s

Case Study #4: Rand Fishkin’s Blog
Our next test was related to domain authority so we picked a hard one. Rand Fishkin agreed to a hijack attempt so we set up a page in a similar way to previous experiments with a few minor edits (rel/prev, authorship, canonical). Given that a considerable amount of code was changed I did not expect this particular experiment to succeed to full extent.

We did manage to hijack Rand’s search result for both [13]his name and [14]one of his articles, but only for Australian searches:
Notice that the top result is our test domain, only a few days old. Same goes for the test blog post which now replaces the original site in Australian search results:
This "geo-locking" could be happening at least two reasons:

1. .au domain hosts the copy
2. .au domain links pointing towards the copied page

Not a Full Hijack

What we failed to achieve was to completely replace his URL in Google’s index (where info: shows our subdomain) which is what happened with Rob’s page. This could be partly due to the fact that the code was slightly different than the original and possibly due to Rand’s authorship link which we left intact for a while (now removed for further testing). Naturally Rand’s blog also has more social signals and inbound links than our previous test pages.

Interesting Observation

When a duplicate page is created and merged into a main “canonical” document version it will display it’s PageRank, cache, links, info but in Rand’s case also +1’s. Yes, even +1’s. For example if you +1 a designated duplicate, the selected main version will receive the +1’s. Similarly if you +1 the selected main URL the change in +1’s will immediately reflect on any recognised copies.


When a copy receives higher PageRank however, and the switch takes place, all links and social signals will be re-assigned to the "winning" version. So far we have two variants of this. In case of a full hijack, we’re
seeing no +’s for the removed version and all +’s for the winning document, borderline cases seems to show +’s for both documents. Note that this could also be due to code/authorship markup on the page itself.

We’re currently investigating the cause for this behavior.

Preventative Measures

Further testing is needed to confirm the most efficient way for webmasters to defend against the result/document hijacking by stronger, more authoritative pages.

Canonicalisation

Most websites will simply mirror your content or scrape a substantial amount of it from your site. This is typically done on the code level (particularly if automated). This means that presence of properly set rel=”canonical” (full URL) ensures that Google knows which document is the canonical version. Google takes rel=”canonical” as a hint and not an absolute directive so it could still happen that the URL replacement happens in search results, even if you canonicalise your pages.

There is a way to protect your documents too (e.g. PDF) through use of [16]http header canonicalisation:

GET /white-paper.pdf HTTP/1.1
Host: www.example.com
(...rest of HTTP request headers...)
HTTP/1.1 200 OK
Content-Type: application/pdf
Link: <http://www.example.com/white-paper.html>; rel=”canonical”
Content-Length: 785710
Authorship

I am not entirely convinced that authorship will do much to prevent search result swap from a more juiced URL, however it could be a contributing factor or a signal and it doesn’t hurt to have it implemented regardless.

Internal Links

Using full URLs to reference to your home page and other pages on your site means that if somebody scrapes your content they will automatically link to your page passing PageRank to it. This of course doesn’t help if they edit the page to set the URL path to their own domain.

Content Monitoring

By using services such as CopyScape or Google Alerts webmasters can listen to references of their brand and content segments online, and as they happen. Acting quickly and requesting either removal or a link back /citation back to your site is an option if you notice a high authority domain is replicating your pages.

NOTE: I contacted John Mueller, Daniel Peng and Frank Dabek for comments and advice regarding this article and still waiting to hear from them. Also this was meant to be a draft version (accidentally published) and is missing information about how page hijacking [17]reflects in Google Webmaster Tools.

PART II:

Article titled "[18]Mind-Blowing Hack for Competitive Link Research" explains how the above mentioned allows webmasters to see somebody else’s links in their Google Webmaster Tools.

4.10.2 Mind-Blowing Hack for Competitive Link Research (2012-11-14 20:37) - admin

Last week I presented a method for [1]search result hijacking. The story got much coverage in the SEO community, perhaps due to the fact that Rand Fishkin’s authority pages were also compromised as part of our experiment. One thing I did not elaborate on in the original article was the peculiar way Google Webmaster Tools handle document canonicalisation.

TL;DR

You can see somebody else’s links in your Google Webmaster Tools as if you were the authorised user of that site. The process involves creation of identical document copy and the results are visible in about two weeks.

In the following screenshot we see what Google calls "an intermediate" link. An old link which points to our old domain 301’s to our new domain. Nothing unusual about this.
There are other instances of the "intermediate link" in Google Webmaster Tools. One of them is related to document canonicalisation process described in a paper called "[2]Large-scale Incremental Processing Using Distributed Transactions and Notifications" by [3]Daniel Peng and [4]Frank Dabek. This is exactly the same process I used in the [5]result hijack article and the most interesting thing is that it works in reverse! (I'll get to that later).

Here's an example of one such case, some of you may remember this website from my hijacking experiment:

As you can see the "intermediate link" notification suggests that my page above receives a link from [6]Rob's website but the thing is, it doesn't. So what's going on?

Well, the page I created is a replica of the original page on [7]http://www.marketbizz.nl/en/ . Google has seen my duplicate version and 'merged' it so to speak. Now I not only see the links pointing to my own version but the aggregate list of links for the one canonical/merged document across all applicable domains. I am seeing the same links the owner of that site would see in their Google Webmaster Tools.

Here's the interesting part, it works in reverse! You don't even have to hijack the result for this to work, you can see the results with the 'loser' URL. If you create a duplicate page with lower PageRank (not very difficult to do that is it?) of any page on the web, you will be able to see its links in your own Google Webmaster Tools.

To test this concept I copied a PDF from another site and simply got it indexed, in about ten days I saw all its backlinks in webmaster tools, here it is: Domains Links doc-txt.com 38 documbase.com 19 pdfqueen.com 7 cmu.edu 6 msra.cn 6 quora.com 5 130.203.0.133 5 google.com 4 berbatek.com 3 podpdf.com 3.blogspot.com 3 seobythesea.com 3 writingsseo.com 3 newyorklastminutetravel.com 2 automotivedigitalmarketing.com 2 123people.com 2 blumenthals.com 2 psu.edu 1 pitt.edu 1 65.54.0.113 1 52opencourse.com 1 chatmeter.com 1 journalogy.net 1 vebidoo.de 1 c4ads.org 1 ryanmc.com 1 ebookbrowse.com 1 uic.edu 1 rightnow.com 1 osti.gov 1 informationweek.com 1 christopherpotts.net 1 sdsu.edu 1 keywordspy.com 1 psugeo.org 1 tagwalk.com 1 delib.net 1

@2013 dejanseo.com.au/blog
Now I can take any page/document from my competitor, place it on a domain of my choice, have it indexed and in a few weeks I’m able to see all their backlinks within Google Webmaster Tools.

It took me exactly 14 days to see the link data of Rob’s website.

Whoa indeed.
I’m pleased to announce that we’ve commenced another search experiment. We’re investigating the impact of link rejuvenation on page’s performance in search. More specifically, we’ll be changing the format of existing links to create new ones.

Background

When Google re-crawls a page and runs into a link which is slightly changed from the last crawl it will treat it as a new link*. What could possibly impact this ‘perception’ includes changes in:
- URL variant
- Anchor Text
- Attributes & Classes
- Location in the Source Code (less likely)

During this process, the page naturally loses one link and gains another. It is not known whether the loss occurs at the same time as the gain so this is one element we’ll be looking out for, although, this will be tricky to isolate.

Secondly, we know that fresh links of fresh documents impact the QDF type scenario but we’ll now be observing the impact of discovery of fresh links on old documents and see if it possibly has a similar effect. Feedback, comments and suggestions welcome. I would love to hear if anyone has experimented with this in
the past.
Confirmed by Google.

4.11.2 PageRank Split Experiment (2012-12-12 01:24) - admin

One of the commonly circulated SEO theories is that links from pages with many outgoing links are not worth much. Dejan Labs team put this theory to a test in hope to find out what really happens.

The Experiment Setup

Number of domains involved: 3 (2 test domains referred to as "A" and "B" and a buffer domain used to bridge PageRank)
Test domain characteristics: Very similar domain format with slight variation.
TLD: .com
Content: Minimal, unique, slight variation on both.
Control Period: 31 July 2012 - 20 August 2012 (used to test any natural fluctuations)
Test Period: 21 August 2012 - 11 September 2012

Key Activities

1. Secure a link on a PageRank 7 sub-page on debian.org for our buffer site
2. Link the buffer site to the test domain marked "B"
3. Observed ranking fluctuations (same location, browser, settings, user, machine) looking for any rouge links unintended and undesirable for the experiment (none were discovered).

Elements of Interest

The page on debian.org contained 4,225 external followed links. Despite very high PageRank, the split was too great for our target page (on the buffer site) to receive any visible PageRank, even after the public TBPR
Experiment Results

One week after PageRank was passed to our target page, the test site rapidly gained position for the selected search term, reaching the highest position of #3. Currently (three months later) the page ranks #2 and holds its position without any variation. The leading result is a PageRank 4 page on a high authority domain. We do not expect these results to change.

TL;DR

Is it worth getting a link from an authoritative domain and/or high PageRank page which has thousands of outgoing links on it already? Yes.

Raw Data:

Tue, 31 Jul at 7:04pm
Website A #64
Website B #65

Wed, 1 Aug at 8:22pm
Website B #70
Website A #71

Thu, 2 Aug at 7:43pm
Website A #71
Website B #73

Mon, 6 Aug at 4:25pm
Website A #171
Website B #179

Tue, 7 Aug at 9:54pm
Website A #69
Website B #71

Thu, 9 Aug at 6:56pm
Website A #66
Website B #71

Fri, 10 Aug at 4:35pm
Website A #64
Website B #69

Thu, 16 Aug at 9:32pm
Website A #116
Website B #115

Sat, 18 Aug at 7:28pm
Website A #115
Website B #76

Mon, 20 Aug at 4:58pm
Website A #121
Website B #79

Tue, 21 Aug at 7:08pm
Website A #113
Website B #287

Tue, 21 Aug at 8:11pm
Website B linked from a site which is linked from http://www.debian.org (sub-page with 4,225 external followed links)

Wed, 22 Aug at 7:01pm
Website A #124
Website B #79

Fri, 24 Aug at 6:44pm
Website A #124
Website B #79
4.11. December

Tue, 28 Aug at 7:07pm
Website A #82
Website B #159

Wed, 29 Aug at 8:42pm
Website A #85
Website B #49

Thu, 30 Aug at 5:38pm
Website A #86
Website B #49

Thu, 6 Sep at 9:31pm
Website A #84
Website B #12

Mon, 10 Sep at 7:45pm
Website A #168
Website B #3

Tue, 11 Sep at 7:56pm
Website A #186
Website B #7