Personalisation, localisation, semantic search: do they work?

Karen Blakeman, RBA Information Services, Reading UK, Karen.blakeman@rba.co.uk

INFORUM 2011: 17th Conference on Professional Information Resources
Prague, May 24-26, 2011

Abstract

Regional and world-wide search engines are battling for market share. Localisation, personalisation and the inclusion of social media are now part of Google and Bing. Both are also implementing aspects of semantic search. But does any of this really work or produce better results? Are any of us who are outside of North America benefitting from these developments or should we be concentrating on regional search tools, specialist search engines and databases? There are some who feel that Google, in particular, is out of control and has gone too far by unilaterally deciding what the user really meant to search for and by automatically including the searcher's social media connections without consent. This presentation will look at what Google and Bing are implementing now and in the near future, with real life examples, and will discuss the implications for researchers. Do we know how much information search engines and web sites have about us with respect to our online connections and search patterns? How does this affect our results? How can we regain control or is it even possible?

Disclaimer

As the search engines now personalise the search process to such a high degree as well as continually conduct experiments with search and ranking algorithms, you may not be able to replicate the results and examples given below on your own computer. Results for any given search can vary from day to day or from one minute to the next. Most of the examples I have given below are from Google because it is the most open of the search engines in explaining what it does, it experiments the most, it can present the most problems when searching and it has infiltrated so much of the web. Other search engines and sites use similar technologies and present the researcher with the same challenge: "Give me what I am asking for, not what you want me to ask for"!

Personalisation is the norm

All of the search engines and many web sites use personalisation, localisation and semantic search in an attempt to offer the user more relevant information and products. Those of us who regularly use Amazon (http://www.amazon.com/) are no longer surprised to see when we log in "New for you.." or "Recommendations for you..." based on your previous purchases. When I buy groceries online with Ocado (http://www.ocado.com/) and proceed to the "checkout" it checks my past orders and asks "Have you run out of...?" A useful reminder that I might have forgotten something.

Perhaps not so welcome is targeted advertising. Every time we run a search and view web pages a multitude of advertising agencies track our every move. Google monitors each search you carry out and which results you click on. The next time you use Google, or a service that uses advertising powered by Google, the advertisements are customised according to your search and viewing history. For example, I gave a talk In Trondheim in Norway last year and I used Google to find a hotel for my stay. When I later went on to YouTube and then the Guardian newspaper, large image advertisements for hotels in Trondheim kept appearing. This continued for several days. Google does give you the option

to control targeted advertising. Go to http://www.google.com/ads/preferences and it will show you the categories of advertising in which it thinks you are interested. You can add and remove categories or opt out altogether. If you opt out you will still see advertisements but they will not be based on your web viewing history. This information is stored in a cookie on your computer and is browser specific. For other advertising networks the Network Advertising Initiative (http://www.networkadvertising.org/) enables you to opt out of over 70 networks.

Facebook thinks it knows best

Personalisation may be an advantage when using a shopping site but is of serious concern when used by social networks and search engines without your knowledge. It can restrict and bias your view of the data. Facebook recently decided to limit the news and information that you see to only those people and organisations you interact with the most. This move was not widely publicised and it was only when people started to ask why some of their friends had stopped updating their status that the change was made clear. You had to go the bottom of your home page to change the default for "Show pages from: friends and pages you interact with the most" to "All of your friends and pages".

The problem here is that there are a number of reasons why you do not interact with a "friend" or a page but **are** still interested in what they have to say. I follow a number of organisations who provide services in my home town of Reading, for example Reading Buses and Reading Borough Council. I follow them because I want to be kept up to date with news and changes to their services and not because I want to engage in a conversation with them. I follow company pages because as part of competitive intelligence. I follow people who have a completely different opinion from me on issues such as energy sustainability and I follow them because I am interested in their arguments; I may or may not decide to discuss their views with them.

Limiting information and search results to those that you "talk" to most is bad practice when it comes to research. One needs to consider all sides of an argument but increasingly we have little choice or control over what the search engines decide is best for us.

How the search engines do it

The results that the search engines give you depend on:

- country version of the search engine used
- location within the country
- language used for the search interface
- browser, version of browser, operating system
- whether you are using a pc, smart phone, tablet
- whether or not you are logged in to a search engine account
- web and search history, black lists, white lists
- the type of search for example person, company, current news, scientific or technical query
- search engine experiments (especially Google)

An identical search run by several people at the same time but on different devices and in different locations will come up with different results. This presents problems for those of us who help and train others on effective search strategies. What appears on your screen may not be what is appearing on theirs. We can no longer advise that a particular search strategy is the best approach to specific type of search – there are far too many variables being considered by the search engine.

The evolution of Google is neatly summarised in a diagram “Google's Collateral Damage: SEO Cat & Mouse Game” (http://www.seobook.com/learn-seo/collateral-damage.php). Google became popular so quickly because the main component of its ranking algorithms was the Page Rank: who is linking to whom and what authority do those links have? The results that Google presented were more relevant than the existing search engines and so Google gained in popularity. Since their launch in 1998 Google has added more search features and the number of “signals” included in the way the results are ranked have increased.

All of the search engines have had to contend with web sites that use the ranking algorithms to their advantage by designing pages with little content or added value yet still manage to appear at or near the top of search results. The search engines continually endeavour to counter this, the most recent and widely publicised changes being Google’s Panda update (Official Google Blog: Finding more high-quality sites in search http://googleblog.blogspot.com/2011/02/finding-more-high-quality-sites-in.html). With so many interlinking algorithms even small changes can have an unexpected negative impact on the ranking of a good quality web site. To override this effect both Bing and Google have white lists that override these algorithms (Google, Bing have white lists of sites not to be Impacted by algo changes http://blog.searchenginewatch.com/110310-175043).

The search engines remain reticent with respect to the details of the signals used to rank pages. Bing has said that it uses 1000 whereas Google claims to use over 200 each of which may have over 50 variations bringing its total to about 10,000 (Dear Bing, We Have 10,000 Ranking Signals to Your 1,000. Love, Google http://searchengineland.com/bing-10000-ranking-signals-google-55473). This sounds impressive but of greater importance is the relevance of the results that these algorithms deliver. As more signals come into play and more adjustments have to be made to compensate for unwanted effects the algorithms start to behave like intertwined spaghetti. Pull on one piece and the whole pile of pasta ends up on the floor. Run a seemingly straightforward search and the most bizarre results appear.

Where are you?

Type in Bing,com or Google.com into your browser and you are usually taken to a country version of the search engine based on your IP address. Although you are still searching the whole of the world-wide local content and web sites are given priority. Both Bing and Google now go even further and encourage you to specify your town or city so that results can be localised even further. This encourages lazy searching, for example if you are looking for somewhere to eat in Prague and your city has been set to Prague you only need to search on restaurants and straightaway a list of possibilities and their location are displayed on a map. This is not so helpful, though, if you have been asked to research the distribution of McDonalds across the whole of the Czech Republic.

Localisation can be used to advantage when researching an industry, person, company or services in a particular country or city. One can choose to go to a specific country version of the search engine and change town or city as required.

The local content that is provided by the country versions of the search engines is sometimes in the language of that country. Google’s ‘Translated foreign pages’ offers the searcher a quick and easy way to search that content. The option can be found towards the bottom of the menu on the left hand side of your results page. You can choose or add a language to the list that is presented to you but Google first offers the language it thinks best fits your query. For example, I am interested in finding out more about the imminent birth of a hippo at Prague zoo. News from Prague Zoo may not be translated from the Czech into English but when I click on ‘Translate foreign pages’ Google automatically translates my search into Czech and translates the pages it finds back into English for me. For information

on Sputnik I am offered Russian and for Edvard Munch Google suggests Norwegian. Take note, though, that this is machine translation and although it has greatly improved over the last 2-3 years it is still far from perfect.

Type of search

The search engines try and determine the context of your search and type of information you are looking for. If Google thinks you are looking for a person it will give priority to social media profiles (Flickr, Twitter, Facebook, LinkedIn etc.). If the topic is a major news story additional pages are added to the results that might not otherwise appear; for example after the Japanese earthquake and tsunami links to the Pacific tsunami warning centre and the Japanese quake person finder were added to the top of the results. Type in the name of a food and Google sometimes displays its recipe results page that includes options for choosing ingredients. If it is a technical or scientific query, Google emphasises papers from Google Scholar.

Very often Google gets it right but I may, for example, be looking for information on the history of pancakes and not recipes. A year ago, a simple search on just pancakes would yield a mix of results with at least one page in the top 10 that was of interest. We have become accustomed to lazy searching but we now have to consider very carefully what to include in the strategy in order to retrieve meaningful results. For my pancake search I have to compose a more ‘traditional’ strategy such as pancakes origins OR history.

Is Google trying to be too clever?

For a long time Google has been very sympathetic to typographical errors. Rather than return no results Google used to ask “Did you mean...”. Google rarely asks now and runs what it thinks was your intended search with link to your original search under “Search instead for..”. For some searches it will not even provide the alternative. A recent search of mine n an English beer called Hewish Mild is a case in point. In a way, Google did find the right answer by putting the brewery that makes the beer at the top of my results list. However, for the rest Google decided that I really meant Jewish mild (the letters J and H are next to one another on the keyboard) and it did not even offer my original search as an alternative! To force Google to do what I wanted I had to put a + sign before Hewish to force an exact match, and I am having to do this more and more with what should be straightforward searches.

By default Google monitors every search and result that you click on and adjusts the results of future searches according to your previous choices. A few days after my initial search on Hewish mild and after research on other beers, I repeated my search on Hewish mild. This time Google included more results from the brewery. A few pages on Jewish mild remain, though, and I am still not offered the chance the “Search instead for...” .

Is Google Scholar trying to be too clever?

(With thanks to Even Hartmann Flood and Sara Batts for the examples).

Strange thing also happen in Google Scholar but often an exact match in Google’s standard web search is not in Google Scholar. It appears that Google Scholar has and additional set of rules that make searching more time consuming and unreliable.

Query 1: Exploration of the Norne oil field in the North Sea

a) Google Scholar looks for the author Horne as well. Analysis of some of the paper shows that Google is not just assuming a typing error (the H key is above and slightly

to the left of the N key). There is an author called Horne working in the area of North Sea oil field exploration. A plus sign before Norne (+Norne) forces an exact match for Norne.

b) If you switch the interface language to Norwegian and run the same search an exact match search is carried out and there is no “Horne” (Note that we are changing the language for the search and menu options, not searching in Norwegian or using ‘Translate foreign pages’.

c) Change the interface language to Swedish and we are back to norne/horne

Query 2: A search for information on a project called EFET

a) Google Web search does an exact match

b) Google Scholar automatically looks for ‘effective’ and we have to prefix EFET with a ‘+’ to force an exact match

c) Changing the interface language to Norwegian results in an exact match

d) Changing the interface language to Swedish and an author named K Efe, hitherto not mentioned, are highlighted

The results of Google Scholar’s efforts to customise results and attempt semantic search can be worse than Google web search. It often requires a liberal does of plus signs and quote marks around phrases to make both Google and Google scholar do what you want. Somehow we have to pass on these warnings to users of Google Scholar who may assume that Google knows best and if they cannot find a paper in the Google results then it does not exist.

Social media and more customisation

Both Bing and Google now include social media in their results. If you are searching Bing from within your Facebook account Bing will take into account the “likes” of your friends when ranking search results. If you are logged in to a Google account when searching, Google may include and give priority to your social circle gleaned from contacts in your Gmail account, Google Reader, Google Groups, Google Buzz and social networks such as Twitter that you may have mentioned in your Google profile. Not only does it search postings, tweets and web sites owned by your 1st level connections but it also looks at 2nd level connections – that is those that are connected to your 1st level connections but not directly to you. The problem here is that for those of us who conduct research for business or academic reasons our social circle may be mainly personal and totally irrelevant to our research. You can only switch this off by logging out of your account but you can see who Google is including by checking the dashboard on your Google account at http://www.google.com/dashboard

Other customisations being introduced include a ‘+1’ button to “approve” a page, tweet or posting in your results list. You will also be able to block specific sites from your searches. Google says it that at some point it will probably use these as “signals” for everyone, not just you. This is worrying because what I consider to a worthless site in my researches could be regarded as essential by someone running different types of queries. Neither of these two features has yet been made available to everyone and Google may change its mind with respect to using them as universal ranking criteria. Google always monitors reactions to changes in its algorithms, for example people may not click on any of the results of a search because they are irrelevant and try one or more different strategies. If the response is poor then Google rolls back to earlier setoff algorithms.

What can you do?

Sometimes it seems impossible to make Google behave but there are some simple tips that can be passed on to users to ensure they are getting what they ask for and a few advanced tips for the serious researcher.

1) Look very, very, very carefully at your results and at what Google is trying to do to your search. What is highlighted? Do the results make sense? Has Google automatically looked for synonyms and spelling variations without telling you

2) Use plus signs before a term to try and force an exact match or quote marks a round a phrase. Unfortunately Google does sometimes ignore these.

3) Change the order of your terms in the strategy. This can radically change Google's behaviour and your set of results.

4) Repeat one or more of your terms one or more time. Again this can radically change Google's behaviour and your set of results.

5) Include advanced search commands for example filetype:pdf when looking for a scientific paper. Google tends to give up trying to take control when you use advanced search features.

6) Enable or disable web history. Sometimes enabling web history so that Google adjusts results according to your precious queries is a good idea, as seen in the above example on Hewish mild. It can, though, bias your results. The only way to decide is to try it out for yourself and see what works for you.

7) Clear cookies and your browser periodically. This has the advantage that it removes the personalisation that Google has unilaterally imposed on you buit the disadvantage that it also removes your own settings.

8) Use something completely different for example local search engines such as Seznam.cz, and specialist databases

Google tries to be too clever and gets it totally wrong

With special thanks to Arthur Weiss and Susanna Winter for their help with this analysis.

What follows occurred in February 2011 over a period of about 8 days when, it seems, Google was testing out a new set of algorithms. We are not sure if everyone was seeing the same type of odd behaviour in Google. The following searches were conducted in the UK using Google.co.uk.

See "Google decides that coots are really lions" http://www.rba.co.uk/wordpress/2011/02/12/google-decides-that-coots-are-really-lions/ and "Update on coots vs. Lions" http://www.rba.co.uk/wordpress/2011/02/21/update-on-coots-vs-lions/ for further details.

The query

I was walking along the River Thames with some friends when we saw two coots fighting on the water. One of my friends said he thought it was mating behaviour so we decided to use Google to settle the argument. The strategy we used was ‘coots mating behaviour’.

Straightaway Google came up with “Showing results for lions mating behaviour”. There was no “Did you mean...” but there was a link “Search instead for coots mating behaviour” and clicking on this link did take us to what we thought were the correct results. But Google was still insistent we really meant lions and asked “”Did you mean lions mating behaviour”.

Placing a plus sign before coots in the strategy gave us “Showing results for +lions mating behaviour”. Putting the whole search within double quote marks gave “Showing results for lions mating behaviour”.

So how did Google decide that coots are lions? Are all coot queries going to be changed to lions?

‘Coots feeding behaviour’ gave us an exact match.

Changing the order of the terms to ‘mating behaviour coots’ gave an exact match.

Repeating the most important term so that our strategy read ‘coots coots mating behaviour’ gave an exact match.

The search strategy ‘coots mating behaviour coots’ resulted in “Did you mean lions mating behaviour coots”

At this point we decided to try and get rid of the lions from the strategy by adding –lions to the search. Google came back and asked “Did you mean cats mating behaviour –lions” and light started to dawn. We think that Google assumed a typing error - we really meant cats not coots - and then did an automatic synonym search, hence the lions.

Google UK is no longer showing lions or cats instead of coots but at the time of preparing this paper Google.cz, Google.no and Google.se suggest cats instead of coots and Google.de suggests cows.

Why did coots feeding behaviour give an exact match? Perhaps a search query frequency algorithm? Or just spaghetti algorithms? I don’t think anyone knows for sure, least of all Google!