The winning mindset: Effective competitive intelligence research on the internet
Terry Kendrick

Business Information Review 2007; 24; 228
DOI: 10.1177/0266382107084890

The online version of this article can be found at:
http://bir.sagepub.com/cgi/content/abstract/24/4/228
The winning mindset

Effective competitive intelligence research on the internet

Terry Kendrick
University of East Anglia

‘Search engines eat your brain. Use them responsibly.’

Abstract

Suggests that search engines are useful but limited in their application for competitive intelligence searching on the internet, and highlights the importance and effectiveness not just of structured searching but also of creativity. Explains some of the technical limitations of internet searching and suggests conditions in which a competitive intelligence search may be made more effective, pointing out that the value an information professional adds is in having some idea in advance of what they are likely to find. Gives details of what search engines will and will not retrieve, and illustrates how search strategies can be improved through use of the available filtering syntax. Suggests that using Boolean logical operators and other features directly in the search box is likely to produce better results than simply relying on the search engine’s advanced search feature. Concludes by re-emphasizing the need for a creative mindset, building on some structure.

Keywords: business information, creativity, internet, research, search engine, search strategy, searching, structured searching

Introduction

Answering broad business information questions by using internet search engines can provide results which are at any point on the continuum between spectacularly useful and incredibly irrelevant. Many people will start a search for business information by undertaking a general search in their favourite search engine using keywords or subject terms. Sometimes they ‘hit lucky’, sometimes they don’t. This article is all about how to ‘hit lucky’ more often by having the winning mindset for effective business information research on the internet.

In this article I reflect on over ten years as an internet searcher and over twenty years as a business information researcher. I believe that, above all
the advice on effective searching, there are two key points for the business information searcher keen to improve internet search results:

- Search engines are useful, but limited, in their application for business information searching on the internet.
- Structured searching is important but creativity is even more important.

The rest of this article looks at these two key points in detail.

**Search engines are useful, but limited, in their application for business information searching on the internet**

A business information need can be anything from a request which a business information professional should offer a 100% guarantee to answer quickly and correctly (for example the registered office of a UK limited company) through to more difficult requests where the business information professional may, in most cases, have a less than 10% chance of answering satisfactorily via an internet search (for example providing a list of suppliers to a particular company). The second question is answerable but very rarely via the internet.

To understand the boundaries of an internet search in answering requests for business information consider two things: it is only possible to find things on the internet which have been mounted there; search engines can only bring back things for you which their search technology has been able to deal with.

**It is only possible to find things on the internet which have been mounted there**

The first point, that you can only find what is there, seems trivial but it is not. A business information professional should know in advance of a search the type of materials that are likely to have been made available via the internet: subject knowledge is, and perhaps always has been, one of the biggest challenges to business information librarians using the internet (Liu, 1990). Of course, this knowledge can never be totally exhaustive but the searcher should have some idea of what he or she is likely to find: this is part of the value the information professional adds to the information value chain before the end user sees it. Without a reasonable idea of what is likely to be returned the information professional is little, if anything, better than an end user who simply throws speculative keywords into their favourite search engine and hopes for enlightenment.

A search of the internet for business information will be most likely to be successful if at least one of the following conditions is met:

1. *It is likely that the information required may have been created as part of the national planning process of a country.*

Why might your information have been created? If you can imagine a national government collecting the information you require to support national planning in some way then you have a chance of accessing it via the internet. Governments are increasingly seeing the internet as a way to disseminate such information and often you can access electronic versions for free where there may be a substantial charge for the hard copy. Given the importance of export markets for most countries, do not always assume that the best country-specific market research will always be found from within the country. For example, an almost Mintel-level report on the UK wine market is freely available on a United States government website.

In addition to national governments you may find that subject specific international organizations such as the World Health Organisation have the information you require. However, beware of incompatibilities between seemingly similar data (e.g. EUROSTAT and national European Union governments’ information) and ensure that you understand not just what statistics exist but how they are compiled. Government production and sales statistics are, for instance, unlikely to be market size statistics, although they are often taken as a proxy for such, as government will tend to use factory gate prices rather than actual sales. Given the nature of internet pages you may find the guides to how the statistics are compiled are not easily found on the page in front of you.
Whatever the pitfalls and checks you need to do, if you can imagine government collecting the information for its own purposes then you have a good chance of finding your information on the internet.

2. *It is likely that the information you seek has been created as a service to a group who join together to fund the creation of the information they need.*

If you can imagine a group of people or organizations jointly creating the information for their purposes (e.g., oil producers through OPEC) then it is likely that a search engine should be able to help you find the business information you need. National and international trade associations are the major group in this category. In recent times a number of trade associations have used the internet to disseminate information not just to members of their association but also to the general public. Sometimes this is up to date, sometimes two or three years old. However, even the older data can, with care, help you to scope the market. Remember also that you may find that the international trade associations may be more helpful than national trade associations.

3. *It is likely that the information required is so important to a number of people or organizations that a third party has found it profitable to collect and report on this area.*

Market research and company information publishers are key organizations within this category and most of these find the internet an ideal medium for their reports. Many of these are downloadable for a fee so the page by page detail of the reports will almost certainly not be searchable via a search engine, whose role will be to get you to the first or publicity page for the report which will normally be kept behind a password or electronic billing front end. Some market research publishers offer free executive summaries online which can sometimes be all you need for your business information requirement.

4. *It is likely that the information required is valuable to an organization attempting to gain influence in some way.*

Pressure groups (cause or promotional driven) often compile information on industry and commerce which goes beyond what might be found in the traditional statistical and company profiling services. Organizations in this category include the Adam Smith Institute, the Portman Group and Friends of the Earth. Websites from such organizations often include free downloadable reports which contain data specifically created for those reports.

5. *It is likely that the information is topical or controversial enough to have encouraged people to discuss or gossip on the subject.*

When researching a company’s activities it is always worthwhile using the news buttons of your favourite search engine to search the newspapers where the company of interest has its headquarters. A local newspaper often has better contacts and is interested in the human dimension to stories which would rarely interest the national financial press. Here, you will get rumours and comment from individuals not seen in the national press.

You will find that, although there is overlap between the general web search and the news search, they are not the same, if only because the ranking systems of most search engines will ‘lose’ many of the small local newspapers in the depths of the found items. You may need to go direct to the website of the local newspaper, as some have free password protected archives which have ensured that, even though you can search from the website, there will not be copies in the search engine cache.

The internet is a superb medium to get your own point of view across. Social networking and blogging activities have created a set of information resources to help people disseminate or exchange views on a myriad of topics, including many that impact on business. Whilst these need to be treated with care because they cannot be said to be authoritative, they can sometimes provide leads as to what is really going on. Many companies now use blogs as a way to disseminate news and details of new products, often written in a more interesting way than when entered onto the official website.

Bulletin boards for job seekers can provide nuggets of potentially interesting information from disgruntled employees. In addition a search for curriculum vitae from people who have recently worked at the companies being researched may prove interesting as individuals may, in the spirit of showing their contribution, reveal things in their CVs which give the researcher insight into recent happenings within the company being researched. You will be surprised just how many CVs are available on the web and, if nothing else, they often include a name and contact details which might be useful for further investigation. Of course, people have been known to be less than entirely honest on their CVs so this type of information needs to be assessed with care!

If you can imagine someone or some organization having at least one of these five reasons to produce the
information you seek then there is a reasonable chance that the information you require exists and is accessible via a search engine.

**Search engines can only bring back things for you which their search technology has been able to deal with**

As noted earlier, in addition to the limitation of ‘you can only find it if it is there’ there is also the limited ability of the search engine to return what is actually there. Search engines can only bring back things for you which their search technology has been able to deal with. It may be there but ‘invisible’ to search engines.

All information searches via the search boxes of search engines simply interrogate a cache of items which their robots, spiders or other intelligent agents have found whilst cruising around the internet looking for web pages to bring back. Most people now understand that when they put search terms into the search box of their favourite search engines then the search is not done real time across the internet. All you are doing as a searcher is searching a search engine's cache, which is a snapshot of various parts of the web at different times.

This is not too much of a problem if you are dealing with a subject likely to be well covered by basic html web pages and likely to have been found and returned to cache by the search engine. The best search engines are very good at sweeping the general internet. However, if the information you require is within an accessible searchable database, is behind a password protected site, or has a number of other technical features which make it difficult for a search engine to take a copy, then this will not be searchable from a search engine search box as it will not have been returned to cache. This is not simply a matter of subscription versus free databases and password protected sites. Even those which you as a searcher can access freely may not be accessible via a search engine’s cache.

Before moving on to discuss this further it is worth pointing out that the caching process does have some benefits for searchers as well as disbenefits. Cached copies are useful when a website is temporarily unavailable. Another interesting feature is that sometimes you can access materials via the cache which are no longer publicly available. For instance, sometimes materials are put up free for a while and then archived behind a paid for service. If the robot or spider picked such items up and put them into cache then there may be a period when looking at a cached copy will be productive when clicking through to the website will not be.

Returning now to the problem of items which cannot be cached by a search engine: unfortunately in business information research a significant amount of potentially relevant information for the researcher is available via a database or other format that is not search engine robot friendly. Most importantly, the key quality authoritative business information sources are sometimes not searchable via a search engine’s cache. This is in effect a double problem – not only might a search engine not find the answer to your business information query but it might find an answer which looks plausible but does not come from an authoritative site.

A classic example of this double problem is Companies House in the UK. This is the authoritative source of information on companies in the UK but cannot be searched direct from a search engine. You have to know it exists, go to its web pages and find a search box to allow you to interrogate its database. The best a search engine can give you is direct access to the web pages of a company information agency containing the data you want. This is likely to be correct unless some terrible event has occurred to corrupt data between the agent and Companies House but this is not really the point. The point is that the authoritative source of information is not searchable directly. However, if you go direct to the authoritative source you can do the search with authority there. Of course, much of the information held by Companies House is not free but basic company details are.

It is not clear how much business information is ‘invisible’ to search engines but it is certainly a substantial amount. Zillman (2006) notes that the ‘Deep Web’ covers somewhere in the vicinity of 900 billion pages of information located through the World Wide Web, in various files and formats, that the current search engines on the internet either cannot find or have difficulty accessing. Dogpile (2007) notes the existence of 11.5 billion web pages of which 9.4 billion are indexed by search engines. This does not appear to include all the material which is accessible via databases.

Nowadays if you have a large database which you wish to make publicly available it is likely that you
will allow access to your server from your website rather than create flat html web pages which would need regular updating. This decision means that the contents of your database cannot be searched directly from a search engine as the search engine would not be able to take page snapshots back to its cache; the most help a search engine can give is to direct searchers to the page where the search box allows access to the database.

There are a surprisingly large number of databases searchable in this way. These range from government databases such as the UK Companies House and Patent Office databases to private sector credit checking databases and very specific university-created databases of bibliographic material.

In addition to this, the very way search engines take back data to their caches means that, in some cases, there is extra data to be found by going direct to the site and seeing the full item. One example of this is Powerpoint slide sets given by companies at conferences or other gatherings – a potentially good source of business information. Search engines have html copies of these in their caches but if the searcher goes direct to the file on the ultimate website then he or she is able to access the slide set notes pages, which are not copied to a search engine cache and thus are unsearchable from there. These notes pages are private prompts and notes for the presenter and can sometimes include information not intended for the general audience. Sometimes just noting the organizations a company presents to can give you an indicator of the market places which it is either in or trying to develop. At other times the presenter has left notes pages which he or she really should not have left. A warning to you all – if you leave a set of your Powerpoint slide sets with an organization then remember to clean the notes pages!

Even if the business information searcher were to accept the ‘invisible’ loss from his or her set of potential resources, there is the further problem that not all search engines will deliver business information equally. A recent comparative study into the four major search engines (Dogpile, 2007) found that first page results delivered by all four for the same search terms can be very different, with very little overlap between them. The study was undertaken in April 2007 for Dogpile, a meta-search engine which delivers results from Google, Ask, Yahoo and Microsoft Live Search. First page search results on the four search engines overlap by less than 1%, while just 3.6% of organic results for a given search term are the same across the four. At the time of the previous study, in 2005, this figure was 7%.

Very few business information researchers would, it must be said, stick with first page results but the underlying point remains important – search engine results ranking means that some highly relevant items may remain invisible unless you search more than one search engine. To see the importance of this in action go to http://ranking.thumbshots.com and try the searching across different engines. You will probably be surprised by the variation in degree of overlap between the major search engines for many business information searches – often only 30% or so in the first 100 or so hits, sometimes significantly less. To help you search a number of search engines easily use a search engine such as Zuula (www.zuula.com), which will provide you with full results from a number of the best news, images, blog, jobs and general web search engines. Beware of simply using a traditional metasearch engine such as Vivisimo, Ixquick or Dogpile: these are excellent at providing a few highly targeted hits for the general researcher from a range of search engines but are rarely enough for a detailed business information research enquiry.

Quite often what is required may not be in the first 100 or even 1000 hits in the major search engines. Search engines tend to assume, quite reasonably, a ‘general’ searcher and their search algorithms will therefore reflect this (through the number of incoming links, site popularity and a long list of other factors) and will bring back the most general hits first. Often business information research starts from a reasonably good understanding of the topic and the need is for the less popular but revealing items.

Beware that, no matter how many hits a search engine will claim to have found for you, even the best search engines will only give you 1000 of those before you have to amend your search. In Google, to ensure that you get all 1000 put &filter = 0 at the end of the command line response and refresh. If you specifically want to find the less popular hits and perhaps unearth unusual nuggets then it is worth using www.live.com where the searcher has the ability to alter the ‘recency’, ‘popularity’ and ‘closeness of match’ bars in the advanced search function. This will significantly alter the returns for your search. Again, do not look for the one true search, seek to find a range of ways to ask the question. The shift away from ‘expert’ indexing in some circles to community indexing through tagging makes this even more important.
You may need to go to specialist sources such as blogs, news, groups and others. Even if these are in the cache, and often they are not in the general cache, such sources are unlikely to get in the first 1000 hits of a general web search – search engine algorithms are looking for the popular overall vote not the business researcher in particular. Remember that you will only get 1000 hits retrieved. The estimated number of hits is not a guarantee of the quality of search engine results.

Having highlighted the limitations of search engines for business information research on the internet, and the importance of knowing your subject to be able to go direct to key internet sources, this article now turns to discuss the appropriate mindset for searching the internet.

Structured searching is important but creativity is even more important

A traditional online host, such as Dialog, will often have a technical architecture based upon a set of fields and a thesaurus. This means that a human being has categorized the items into groups regardless of the words used in the text. Given this, Boolean searching is very effective as it can precisely bring together items which are very similar to each other. When searching traditional online hosts, traditional library and information cataloging and classification skills will give the searcher the right mindset for effective searching.

It is not quite so simple on the internet. The ‘fields’ that exist are almost accidental and need to be used creatively. In addition, there is no thesaurus underpinning the architecture. It is best to think of the internet as simply a string of text with all the imprecision implicit in that. Boolean searching can be useful but, because of the ambiguities within language, it will certainly not be quite so effective on the internet as when used in traditional online hosts.

Let us take a potential business information requirement and look at a creative, yet structured way to deal with it: ‘What is BP doing in Russia? How is this likely to progress?’ Here is a simple set of search responses to this increasing in complexity or creativity with comment on an appropriate mindset to have. All search engines have slightly different functionality so what works in this example may not work in other search engines. This example uses Google. You will notice that the best responses need your brain and the search engine merely responds to your creativity. Beware that you may get significantly different numbers of results when searching at another time.

Table 1: Search undertaken on Google.com on 2 August 2007

<table>
<thead>
<tr>
<th>Search term /phrase /operator</th>
<th>Number of results estimated by Google to be in its cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1] BP</td>
<td>136,000,000</td>
</tr>
<tr>
<td>[2] BP Russia</td>
<td>2,190,000</td>
</tr>
<tr>
<td>[3] “BP expects” OR “BP intends” OR “BP will” Russia</td>
<td>32,400</td>
</tr>
<tr>
<td>[4] 2008..2025 “BP expects” OR “BP intends” OR “BP will” Russia</td>
<td>57,300</td>
</tr>
</tbody>
</table>

Search [1]: BP.
Estimated results: 136,000,000

Nobody would seriously run such a simple search in response to the question. However, this would seem to be the most inclusive of all elements as all items related to BP and Russia might be expected to be in this set somewhere. The important point about this search is that even though all the relevant items are logically in the set seemingly on offer Google is, in reality, only going to show you significantly less than 1% of the items it estimates that it has on Russia and BP. This is because, even with encouragement, Google will only ever show you 1000 hits from this set of 136,000,000. Items related to current and future operations in Russia are unlikely to have many links or relative popularity and hence will be very low in the ranking, most well lower than position 1000 (the deepest you will be able to go in the search without modifying it).

Search [2]: BP Russia.
Estimated results: 2,190,000

Searching for keywords ‘BP’ and ‘Russia’ will certainly bring back relevant items but has a number of problems in terms of a fully professional search. In particular:

- The search results may well include general business pages which talk about BP in Columbia and a
statesman’s visit to Russia as well as truly relevant items specifically describing BP and its operations in Russia.

- The search results will include everything in the search engines' cache regardless of the date of publication or reference period. The future after 2007 will not be the sole filter for materials.

A traditional online host will usually allow date searching at this point or the use of more sophisticated event codes. There is no clearly defined way to do this across internet search engines and often the only support is to be able to search by the date the item was put into the search engine's cache. Google has a number range command which we will use as a creative proxy for date search later in this search.

A search engine’s ranking algorithm may provide some help here and from what it knows and has learned about the way people tend to search it will put some ranking and order on your results. Nevertheless this is not very precise and you can do better by using your brain to force your search engine of choice to work to its very limits beyond keyword searching. We now turn to this.

Search [3]: “BP expects” OR “BP intends” OR “BP will” Russia. Estimated results: 32,400

As noted earlier, the internet is best thought of as simply a string of text and with this in mind we can go some way towards solving the problem associated with search response [2] – how do we make sure we have items related to the future? By using future words and part sentences we can draw out items related to a future time. In our search [3] whatever comes after ‘expects’, ‘intends’ or ‘will’ (and many other future words you could include as part of your OR search here) will be related to the future. We could of course add other small phrases with forward-looking words and the OR operator. A good place to identify these is from the vocabulary of forward-looking statements which is often explicitly stated in company annual reports.

Search [4]: 2008..2025 “BP expects” OR “BP intends” OR “BP will” Russia. Estimated results: 91,400

Although [3] will return relevant material there will still be a lot of unwanted items in the set of hits. This is because something written in 1999 about 2001 will be included in the search results. Search response [3] can be improved by using the Google number range function and adding 2008..2025, or similar. 2008.. is another option because this will look for all numbers 2008 or greater. This is a number range command (will search for 2008, 2009, 2010, 2011, 2012 all the way to 2025 or even infinity if the command is simply 2008..) but can be used as a proxy for a date command. Note, however, that this will not be entirely precise as anything else that has a number within the range (e.g. model number 2020) will be included in the results even if it is not a date.

The hits will now reflect both BP, Russia and future words but also have a number between 2008 and 2025 (or whatever future ‘date’ you choose) in the text. The searcher might reasonably hope that a lot of these are in fact dates and that the hits reflect articles with comment or predictions about the future of BP in Russia.

Notice that the number of results returned is counter intuitive to what might be expected by adding a further part to a search line [3] (in this case the range command 2008..2025; in the absence of a clear operator, this ‘should’ reduce the number of hits, not increase them. (Google defaults to ‘and’ when parts of a search sit next to each other without a specific logical operator.) Remember that Google is only estimating, not counting, so there will always be anomalies. This effect happens in this search wherever you put the range command within the command line with a difference of only 3–4000 results. This is just one example of the need to be creative with search engines and not look for the one true way to do the perfect search.


Search response [4] should be a quite acceptable response to the enquiry. However, this is only the beginning of the searcher's potential for a creative response. Search [5] is an example of a further developed search. Note that as there are less than 1000 hits resulting from [5] this is the only one of our five searches where you will be able to see all the results.

Google offers an ‘Advanced Search’ mask which prompts the search with options for searching by some of the key ‘fields’ possible in the search. Perhaps the most useful are filetypes and domain (site) searches.
although others are available and have specific uses on specific occasions. Other operators such as the * to replace a missing word and the tilde (~) to look for synonyms are also of regular use in business information searches.

Search [5] utilizes some of these options. By use of the domain command combined with the internet country code for Russia we can add an OR site:ru to our Russia keyword. This expanded command around an OR operator will return hits either with the word ‘Russia’in or from Russian sites. Note that these latter sites may not include the word ‘Russia’ but if from Russian sites they are quite likely to be relevant. In addition search [5] reduces the number of hits by looking for pdf files (likely to be of reasonable quality if someone has bothered to turn them into pdfs) or ppt files (to draw out conference presentations and similar materials).

Note however, that if the searcher had done the above BP search via the Google ‘Advanced Search’ box then search [5] would not have been possible as ‘Advanced Search’ only allows one filetype and one domain to be chosen per search – filetype:ppt OR filetype:pdf could not have been specified. For this reason it is always better to write searches straight into the search box, as a creative business information professional is always more advanced than the ‘Advanced Search’ mask.

The more complex your search gets, the more suspicious Google becomes of your intentions, and when you have combined a number of options and complex logic you may be confronted with a response that ‘your query looks similar to automated requests from a computer virus or spyware application’. At this point you will be required to confirm that you are not such an application. This is simply done by typing a given word into a confirmation box and you may resume your search.

Now of course there are other ways to deal with this query through multiple keyword searches but the example above just serves to illustrate that there are many ways to use a search engine to research a topic. Do not just do one search, do many. Some would search for universal search strategies (e.g. Barsky and Bar-Ilan, 2005) but this is only the beginning of effective search. There are many effective ways to search. Each will bring texture to your search results. Remember that each way you search will deliver results in a different ranking, so you will often have access to different material for each way you search. Try the above BP searches and look at up to the first 100 hits for each. How similar are they? Then try putting the commands in different orders (yet still respecting Boolean logic) and you will find interesting, if often slight, variations which make sense but are too many and varied to include in this article.

Searching the internet needs a creative mindset building upon some structure. An internet searcher needs to be constantly aware of the items which might be searchable and the way in which a favourite search engine will return hits. This article has attempted to build into the searcher’s mindset an understanding of the limitations of the internet for business information research and to stress the need for creativity in internet search rather than to look for the one true way.

Search engines eat your brain. Use them responsibly. Whilst it is good to be a first class technician in your favourite search engine it is much better to be a first class technician in the sources which might answer your enquiry. However, when you are beyond the limits of your subject knowledge and desperately need the help of a general search in a search engine, remember to move beyond simple keywords and subject terms as quickly as possible. Even then be aware that, until the semantic web becomes a reality, the ambiguities of language in keywords and subject terms will inevitably continue to provide frustratingly irrelevant hits amongst the useful.

References


