What's New in Search: Techniques and Technologies

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

INFORUM 2005: 11th Conference on Professional Information Resources
Prague, May 24-26, 2005

Abstract

Over the last year, mergers and acquisitions have resulted in the disappearance of some of the more well known search tools. In their place, though, totally new services have sprung up. Many of these offer unique advanced search features and are contenders for Google's number one position in search. Newer technologies such as RSS, wikis and desktop search are also altering how we look for and manage information. This presentation will identify the best of the new search tools, essential techniques and features in search today, and how they can be used to improve relevance and reduce information overload.

Introduction

Everything changes so quickly when it comes to search tools: just when you think you have caught up with all the new features in your favourite search engine, it comes up with a batch of new ones. Then well known and established services are acquired and disappear whilst completely new players appear out of nowhere. As if that were not enough new technologies emerge to confuse and overload us even further. For many Internet searchers Google is the number one search tool, but there is a plethora of new services springing up and some of the "old dogs" are learning new tricks. Google is being given a run for its money as others compete for Google's crown as "King of Search". This review looks not only at Google's new search features but also at some of the more innovative technologies and approaches to search that are emerging.

Google

http://www.google.com/

Apart from its stock exchange listing, Google seemed to go into hibernation during much of 2004. There was little in the way of new features and it looked as though its database was not being significantly updated. Then, in the autumn of 2004 and in response to the launch of Microsoft's new beta search engine, Google doubled its database to 8 billion pages and a new numeric range search was added to the Advanced Search page. It did not stop there, though. Google then launched Google Print, Google Scholar, Google Libraries, Google Desktop Search and Google Suggests to name but a few!

Google Print

Google Print is aimed at publishers both large and small and books are supplied by the authors and publishers themselves. In the search box on google.com's search page type in 'books about..' followed by your term or phrase and Google lists those that match at the top of the results list. At present, it appears that there is a maximum of three books listed per search but that could just reflect the number of books that have been supplied by publishers. Google displays an image of the page in the book that mentions your terms the most, and you can view a maximum of two pages either side. There are also links to bibliographic information, to book stores or the publishers own site where you can buy the book.

The Google Print is programme is now being expanded to include selected books from the libraries of Harvard, Stanford, the University of Michigan, and the University of Oxford as well as The New York Public Library.

Google Scholar

Google Scholar (http://scholar.google.com/) enables you to search for "scholarly literature including peer-reviewed papers, theses, books, pre-prints, abstracts and technical reports from all broad areas of research." A wide range of academic publishers are covered such as professional societies, pre-print repositories and universities, as well as scholarly articles available across the web. There is no source list so it is down to guesswork and experience in working out what is covered.

The advanced search options are limited to author, journal and date of publication. The author search is unpredictable and results vary depending on the format in which you type the name, for example 'K H Blakeman' and 'KH Blakeman' (without the space between the K and H) give different results.
Google Scholar automatically analyses and extracts citations. This means that your results may include citations of older works, books or other off-line publications. The results are listed by relevance, which is a disappointment for those of us accustomed to sorting this type of literature search by date, author etc. Google Scholar is still in beta so there is hope that at least a date sort will be added.

Google Suggests

Google Suggests can be found on the Google Labs page (http://labs.google.com/). Simply start typing in your search and Google comes up with a list of suggestions for completing your strategy together with the number of results each will give. It is worth visiting Google Labs on a regular basis as this is where the really interesting beta test services can be found.

Alternatives to Google

So what is wrong with Google? In most circumstances nothing. Google comes up with good results most of the time for the majority of users but there are times when one just cannot find relevant information. This can be because Google, for a variety of reasons, has not covered the pages that are relevant but more often the most relevant information is buried way down near the bottom of your list of results. Different search tools rank the results of the same strategy differently and what is number 22,736 in Google's list of pages may be number 5 in Yahoo's.

Try comparing the performance of different search engines in Thumbshots Ranking (http://ranking.thumbshots.com/). This covers Google, Yahoo, AltaVista, MSN, AltheWeb, Teoma and Wisenut, and you can compare two at a time. Type your search strategy into one of the search boxes and then select the tools you want to compare. Thumbshots compares the first 100 pages found by the search engines and represents them as two lines of circles. The pages that are common to both search engines are represented as solid blue circles and linked by a blue line. This is a good way of testing whether it is worth trying your search in another search tool.

Yahoo

http://www.yahoo.com/
http://search.yahoo.com/

Over the last 18 months Yahoo has bought up Inktomi, Overture, AlltheWeb and AltaVista. At the start of 2004 Yahoo stopped using Google as its partner for web search and launched its own web database. Yahoo has not revealed the size of its database but it is certainly giving Google a run for its money and it does sometimes gives better results than Google.

How do Google and Yahoo compare?

1. The advanced search features of Yahoo and Google are very similar.

2. Yahoo has the advantage that it supports all three Boolean operators (AND, OR, NOT) and parentheses for nested searches. Google only supports OR and you cannot carry out sophisticated nested Boolean searches.

3. Google has a link: command that finds pages that link to your specified URL, but it only find links that match your URL **exactly** and shows only a small selection of linking pages. Yahoo's link commands are far more comprehensive. 'Link:' finds pages that links to a specific URL, for example link:http://www.rba.co.uk/search/, whilst 'linkdomain:' finds pages containing links to any page on the specified domain, for example linkdomain:rba.co.uk. The link command is a useful way of finding similar types of pages, the assumption being that pages that link to another are often similar in content.

4. Google indexes the first 100K of a page: Yahoo indexes up to 500K. This may not sound significant when searching standard web pages but it can make a difference when looking for file formats such as PDFs, which can be very large.As well as the usual PDF and Microsoft Office file formats

5. Yahoo allows you to restrict your search to RSS/XML formats (Advance Search screen under filetype).

This is ideal if you are searching for news feeds on a particular subject.

6. Google has a number range search.: Yahoo does not. In Google, specify two numbers separated by two full stops with no spaces, and include a unit of measure or some other indicator of what the number range represents. For example DVD player £50..150, or toblerone 1..5 kg. Number ranges can be years, weights, prices, temperature etc.

7. Google has a synonym search: Yahoo does not. Precede your term with a tilde (~) and Google will look for similar or related terms.

8. Both Yahoo and Google have an excellent current news search with similar advanced search features, including options to limit your search to one source or to sources from one country. I have found that for most of my own searches Google sometimes has better International coverage, but it does vary depending on the search. Both offer news alerts by email but Yahoo can also provide them as RSS feeds (available only in Yahoo.com at present).

MSN

http://www.msn.com/

Microsoft launched its new web search engine in autumn 2004. It has a simple uncluttered default search screen but I found that results for my test searches were poor when compared with Google and Yahoo. Advanced search features are under "Build a search" and are restricted to the link command, country/region, language, and site/domain. There is no file format command. The News search offers RSS feeds for alerts, which are provided by Moreover.

Exalead

http://www.exalead.com/

Exalead was launched in October 2004. With 1 billion Web pages in its index it may seem small when compared with the other major players, and in particular with Google who claims to have 8 billion pages, but the quality of Exalead is excellent and there are some superb search features.

To begin with it supports wild cards - an asterisk - that can be used to represent any number of letters after a specified string of letters. Wild cards, or truncation, were a standard feature in most search engines but for some reason it was abandoned by them two to three years ago.

Then there is word stemming that can be set as the default under preferences, and which finds the variants of a word such as plural/singular and verb conjugations.

If you are not sure how to spell a word try phonetic search or the approximate spelling search on the Advanced Search screen. Another approach is to use the pattern matching feature.

Full Boolean search is supported (AND, OR, NOT) and there is a proximity command (NEAR) that will search for words within 16 terms of one another. Filetype, country and language search options are all supported, and RSS and blog search options are promised for the near future.

When it comes to the results the usual list of pages is augmented with a snapshot of each web page (you can switch this off if you prefer or just display the snapshots), suggestions for related terms, options to view the results by geographical location and file type (for example .doc, .pdf, .xls).

Find.com “True Business Search”

Find.com was launched in June 2004 and concentrates on business search. It combines results from its own index of business web sites with those from some of the major search engines. It also finds priced reports - "premium research content" - from sources such as Thomson Gale, The Gallup Organization, Frost & Sullivan, BNET, Marketresearch.com, Datamonitor etc.

The Web option is a meta search tool and appears to cover at least Google, Yahoo, MSN, AltaVista, AlltheWeb, SmealSearch, Scrius and About. This option also includes Business Web, which are business web sites identified and indexed by Find.com.

Find.com automatically searches for all of your words but supports the standard Boolean AND, OR, and NOT operators all of which must be in capital letters. The standard double quotes can be used to specify phrases.

The Advanced Search screen has options for searching the full text (the default) or titles only, and keyword search (default) or concept. The concept search takes your terms and looks for similar or related terms. For example, type in automotive and Find.com will pick up references to automobile, motor trade, motoring etc.

There are several more advanced features but these only apply to the Business Web and Premium Research collections. You can use a question mark (?) inside a word to represent a single character, or an asterisk (*) inside or at the end of a word to represent one or more characters. If you are not sure of spelling use the “fuzzy search” feature by adding a tilde (~) at the end of a single word. There is also a proximity option. Place your terms inside double quotes and then a tilde at the end of phrase together with a number. For example: "research development"~5 looks for the terms research and development within five words of each other. You can also boost the relevance of a term or phrase in your search by using the caret, "^", symbol together with a boost factor (a number) at the end of the term. For example: “production statistics”^4

Answers.com

http://www.answers.com/

GuruNet launched its free reference service Answers.com at the beginning of 2005 offering "the best definitions and explanations for over 1 million topics". Information on people, places, words and names is drawn from dictionaries, encyclopaedias and selected web sites. A search on Winnie the Pooh, one of my regular test searches, came up with information gleaned from Who2, the New Dictionary of Cultural Literacy, and Wikipedia. For the artificial sweetener aspartame Answers.com came up with definitions from a standard and a medical dictionary, an entry from Wordnet, the entry in Wikipedia including its Chemical Abstracts Number and structure, and translations into various languages. There are links to other topics that mention your term or phrase and to searches on Google for web pages, images and news.

Wikipedia

http://www.wikipedia.org/

Wikipedia is not new but it is only now that it is starting to gain respectability as a reference source. Wikipedia uses wiki technology and is a free encyclopaedia that anyone can edit. And "anyone" really does mean anyone! That may sound like a recipe for disaster and in fact some pages, often those about politicians, are regularly "vandalised" but they are usually restored and corrected quite quickly. The guidelines on authoring an article require that an unbiased view is presented so if you are researching a subject there are always links to alternative opinions that you can follow. Do not be deterred by the openness of the technology - the authors who take the time to write the articles are very knowledgeable in their subject area and monitor their pages for inappropriate editing.

Turboscout

http://www.turboscout.com/

This is not a search tool in itself but an interface that allows you to search the results of 23 search engines by typing in your search, and then clicking one by one on the engines that interest you. This may sound tedious but the really useful aspect of it is that **you** decide which search tool to use. It could be two or three under the Web tab for example Google, Yahoo, Vivisimo and then you could switch to the Reference tab and look in Wikipedia, FindArticles or Scirus. This is a very useful tool for quickly running your search across selected alternative search tools.

Multimedia

All of the major search engines offer some form of image, video and audio search. Image search options have been in existence for several years but the current battle ground is for audio and video/TV search. Blinkx.tv is worth a special mention because it attempts to index the **content** of audio broadcasts whilst other tools index only the text descriptions of the programmes. The transcriptions of radio programmes often look odd because they are computer generated and the computer does not always get it right. Nevertheless,

Blinkx.tv is well worth trying of you are trying to track down audio broadcasts on a topic or those that mention a personality.

RSS

RSS has been around for several years but is only now being adopted as a means of receiving news and alerts. RSS stands for Rich Site Summary, RDF Site Summary or Really Simple Syndication. It is a way of delivering news, headlines and alerts direct to you the reader. RSS feeds are designed to be read by RSS programs. The feeds are gathered together in a single location in your feed reading program and do not clutter up your email inbox with dozens of separate messages. A good RSS reader will allow you to specify how often your feeds are updated and how long headlines are to be archived. You can specify how often your alerts and feeds are updated and they can be easily deleted when you no longer need them. Examples of RSS readers are Feeddemon (http://www.feeddemon.com), Feed Reader (http://www.feedreader.com/) and Bloglines (http://www.bloglines.com/)

Most of the major news services offer RSS versions of their services - just look for the RSS or XML logo - and both Yahoo and MSN offer RSS feeds for news alerts. If you are looking for an RSS feed on a particular subject the Yahoo Web search advance search screen has a filetype option that allows you to limit your search to RSS feeds

Desktop Search

When Microsoft announced that it was working on a tool that would combine Internet with desktop search, the competition made it clear that they were going to try and beat them to it. The idea behind desktop search is that as well as, or instead of, scouring the web for information on a topic you can also search documents stored locally on your computer. These “documents” can be cached web pages, email messages, Microsoft Office documents, PDFs. They do not search every file format, though, and are no replacement for well organised local document folders.

Copernic Desktop
http://www.copernic.com/
Copernic is well known for its search agents that cover multiple search tools and "invisible web" resources, as well as for its web page tracking program. By default it indexes the My Documents folder but you can customise it by adding other directories or even whole drives. Copernic Desktop can search PDF, XLS, PPT, DOC, RTF, TXT, WP and HTML files. It also indexes your search history, Internet Explorer favorites, Outlook email and contact lists. The most recent version has added Mozilla and Firefox support.

Google Desktop Search
http://desktop.google.com/

The first beta version of Google Desktop Search (GDS) gave the impression that it was launched before it was ready in an attempt to beat Microsoft and Yahoo. It was limited to Microsoft Office formats, Outlook, Outlook Express, AOL Instant Messenger, TXT and HTML but did not index PDFs. The most recent version has added support for PDF files, Mozilla, Firefox, Netscape and Thunderbird. There is also a plugin that indexes Open Office and Star Office file formats.

Some of the original security and confidentiality issues have been addressed but you need to be aware of their significance and know how to switch the options off. GDS can index cache secure https pages that you view, for example online bank statements: untick the relevant box under Preferences. Password protected and encrypted files can also be indexed: again untick the box under Preferences.

Another problem is that documents remain in the GDS cache on your PC even after you have deleted the original files. So it is possible for anyone using your computer to unearth those embarrassing emails and online chats that you thought were long gone. More importantly for corporate users, though, is that this feature could conflict with document retention and management policies. There are instructions for removing documents from the cache at http://desktop.google.com/features.html#remove but it is far from straightforward and you have to know exactly which files and emails you want to delete.

MSN Desktop Search

http://www.msn.com/

MSN Desktop indexes Office documents, Outlook and Outlook Express email, photos, music and email attachments. You can specify exactly what you want it to index: email and My Documents, email and all of your hard disc, or individual directories. The main drawback with MSN Desktop is the lack of support for non Microsoft email clients, browser and document formats.

Yahoo Desktop
http://desktop.yahoo.com/
Yahoo is the latest search engine to launch a desktop search program and, in my view, it is by far the best so far. It searches the usual “popular” file formats such as PDF, Microsoft Office, html etc "plus over 200 more". A full list can be found at http://desktop.yahoo.com/filetypes. It retrieves any file on your computer and displays results as you type. The preview option is excellent and most files look exactly as they should in terms of layout.

What next?

Who knows! Google will continue to be **the** search engine that many people go to first but Yahoo offers strong competition. Both continue to add features but it is a question of how relevant these are to us. For example when Google launched its weather search, it decided that only people in the US were interested in whether or not it was going to snow or rain during the coming week! New tools will continue to be launched. Many of these will disappear without trace. Others will fail to live up to expectations, and features that promise to be “coming soon” will never materialise.

If I had to pick one to watch, it would have to be Exalead. It is a comparative lightweight in terms of database size but has some seriously good search options. If it continues to develop new features and its Desktop Search as promised I believe it could be a serious contender for Google's crown.