Keyword searches and labelling

The relevance of it all

Kiron Kasbekar

Words are not enough to say how thankful we should be to search engines, which have made it so much easier to find information than ever before. They search through billions of Web pages to get us a choice of results we can sift through to find the information we need.

Having said that, we need to move on. For business executives searching for business information on the Web, scouring the Web has become a scourge.

Analysts have found that in a typical mid- to large-size company, **millions of dollars are wasted each year** trying to find critical information. According to a Washington Post Survey, 17 per cent of decision-makers spend 5 hours per workday and 7 per cent of them spend 5 hours per weekend day on the Internet.

Consider this: how often do we really need to search billions of pages? More often than not, the information we seek resides in a handful of websites. But since a search engine-based search is so wide-ranging, the relevant results get mixed up with a lot of irrelevant ones.

If we had been paying for these searches in cash, we would have said we were doing an over-kill. Since we don't have to pay, nobody bothers about the cost. But there is a cost to pay.

Wasting company time

We end up wasting a lot of time checking through many search results that seem relevant but actually are not. Going through even the first few pages of the results can take an entire afternoon.

The problem is compounded by human nature. We never stick to the straight and narrow. More often than not we get sidetracked by some interesting new information that has nothing to do with our original purpose. We may acquire a more holistic view of many things through such searching and browsing, but we may have become inefficient executives in the process.

Some search engines tell you they will arrange the results by order of relevance. Some searches list the results by 'relevance', even going to the extent of giving percentage score of relevance. When you actually open the document, you wonder how on earth the score was given. The truth is that relevance is often determined in a mechanical way; for example, by the number of times your keywords occur in the 'description' of the document.

The trouble is that this by no means assures that the main content of the document is relevant.

That is because software programmes are not yet intelligent enough to judge relevance. You need expert systems to judge relevance. And there is nothing yet to beat the expert human touch.

The ideal situation

My proposition is that if we want a really relevant search, we should not search an ever-growing volume of pages; instead, we should search only a relatively small number of relevant websites. If the base of the search (the list of websites to be searched) has greater relevance, the results of the search will have greater relevance.

The ideal situation would be to need to look at a single website for your entire information requirement. If you are in the business of, say, making cars, wouldn't it be great if all your information requirements were met by a website that provided news on all the cars in the world, all the materials and components that go into the car, the companies that make all these things, the dealers who sell the vehicles, on car marketing and logistics, car finance, capital market trends related to car and auto component companies, recruitment and salary trends in the industry in different countries, and so on?

Unfortunately, we do not live in an ideal world. But if we cannot find a single website that meets all our requirements, do we need to go to the other extreme and search a million websites? The practical thing to do is to find a number that is somewhere in between — and closer to one than to a million. Maybe a hundred websites? Two hundred? A thousand?

You gain nothing from duplication, except verification. When you are doing serious research, reading the same news in two or three different sources helps to identify discrepancies and mistakes. Other than that, going through multiple versions of the same information merely wastes management time and money.

The key then is to optimise on a number to ensure that

Keyword searches and labelling

you do not miss out on anything important. You need to select websites for their relevance to your requirements rather than by criteria like 'popularity', used by search engines. I am not saying that search engines are wrong in using popularity as an important criterion for sorting search results. Popularity acts as a proxy for credibility. If more people use, say, Reuters or Bloomberg for news than, say, a local city publication, then the users of the search engines are more likely to find the search engine ranking more satisfactory.

This applies to search engine users, who are an extremely diverse lot. The approach for business users must be different. They are in a position to decide the relevance of websites.

Count on experience

How do we decide which are the most relevant websites to check out? You need neither rocket science nor artificial intelligence to answer that question. The answer is simple: **experience**.

Experience tells each of us - at least those of us who are accustomed to doing research — where to search when we seek certain kinds of information. We may simply check out the Reuters, Bloomberg or Financial Times websites if we are looking for some recent business news (and The Wall Street Journal if we are paid subscribers to it). Or we may open the BBC, CNN or Guardian websites if we are looking for political news.

Now that may sound very limiting. But it's not always so. You will often find, especially with breaking news, that most of the obvious sites are carrying more or less the same information, based on a news release or disclosure.

If you need to delve deeper, look for background, read about trends, analysis and comments, you may need to expand your search through more websites, including more media sites, corporate sites of companies in and related to your industry, some stock exchange sites, maybe some regulatory agency sites. But you still do not need to run through billions of pages of information, including football club, and travel and tourism websites.

It might be argued that an individual's experience is limited - and that's why an impersonal keyword-based search across the Web is superior. The correct answer to the individual's limitations is not a Web-wide search; it to make the broad selection of websites to search on the basis of the experience of many people. And users should be allowed to keep adding to the list of sites to be searched.

Proper approach

It is unlikely that we will find a single, simple and unique product that will solve our information search needs. But we can certainly adopt the right approach to how we look for, store and retrieve information.

The **order of search (and related management)** should be this:

1. Routinely get news and other information feeds that get downloaded either as the most recent full content where possible or as links to the latest information directly to your server (either a web server or a local network server). Ensure that all the latest information from the desired sites is downloaded. Routinising and automating this process will straightaway save a massive amount of time and salary costs.

2. Allow labelling of the downloaded content and links by a central librarian or knowledge officer. Allow every user to add his or her personal labels to the documents and links. Personal labelling brings search in line with subjective, personal preferences. Three people may label a single article in three or more different ways - one may attach the 'marketing' label to it, another may attach 'marketing' and 'people', and the third may label it 'notes for next week's strategy conference'. Different people see different subjective relevance in content, which a mechanical keyword search can never fathom.

3. Allow users to bookmark documents and links to give them priority status.

4. Then allow users to first search through their book-marked documents and links by using single or multiple labels as search criteria. This means that the search is done on a small but relevant number of documents, and the search results are likely to be most relevant. This will also mean a reduced load on the server and on the network bandwidth.

5. If the search does not yield satisfactory results, let users search through all the downloaded content on the server.

6. If that too does not get the necessary results, go for a search engine search across the Web. Chances are that you will rarely need to do so.

Try labelling

Labelling is a good alternative to inefficient keyword searches across the Web. The question is: Somebody has to create all those labels — how do you create labels efficiently?

Labelling works through databases, which allow query-based searches that are flexible - you can narrow them down or expand them, but the search is disciplined by the way in which the content has been categorised

Keyword searches and labelling

through labels, or other fields.

Labelling requires a certain amount of experience and expertise, but not all that much. You can use external labelling services like Informachine from The Information Company, which is into various knowledge management solutions. Or you may deploy one or two people (or more, depending on the volume of content your company needs to download) to do the job. Or you can use a combination of an external service and your own internal people.

A new look at labelling

How should the labellers go about their job?
The mechanical thing to do is to take every proper and common noun in the text and make it a label. But that is not labelling. That is indexing, which your operating software can do anyway to allow a keyword search through the directory or database.

Labelling must use an understanding of a hierarchy of relevance. To apply this we must distinguish between two types of documents:

1. 'Short' documents, such as news reports and articles in the media, press releases, brochures, case studies, analyst and product presentations, white papers, office memos and invoices or purchase orders.

2. 'Long' documents, such as the World Bank's World Development Report, the Indian government's Economic Survey, government budget documents, corporate annual reports, sustainability reports and suchlike.

Short documents: A quick scan of short documents can tell you what the document is about. Such documents will often have a primary subject and theme, and there will be secondary and tertiary subjects. The primary subject and theme are what determine the greatest relevance of the document; the secondary and tertiary subjects indicate lower levels of relevance. The labelling must be aligned with this order of relevance.

For example, take a news report about the world's biggest steel producer, Mittal Steel, making an unsolicited bid for the world number two, Arcelor. The primary theme is M&A, the primary subjects are steel, Mittal Steel and Arcelor. The report may further discuss how trading in the stocks of these two companies was suspended, what has happened to the shares of other steel companies, and give background information on Mittal's earlier acquisition of LNM Steel and Arcelor's recent acquisition of Dofasco. These are of secondary or tertiary relevance, depending on whether your enterprise is more interested in stock markets or in steel.

You may not need to capture every single proper name in the report as a label. Users can always use a keyword search to find the minutiae.

Long documents: Long documents usually have an executive summary or overview that gives a bird-eye view of the document. It's not hard to decide that the World Development Report is about 'development' and the 'world' or 'all countries'. You can also quickly figure out the secondary labels, such as 'poverty', 'employment', 'literacy', and so on.

Where you really get stumped with long documents is the tertiary labelling. It would be a waste of a librarian or knowledge officer's time to ask them to find and label every single subject, such as individual welfare projects referred to in the report, or the names of towns and villages cited, or the names of people who find mention.

The simple and sensible approach to this problem is to not even try to do such tertiary subject labelling. Just leave it to automated indexing by standard software, such as IIS, and let users find such information through keyword searches.

That, of course, assumes that such long documents have already been downloaded and placed on your server, and you have a document management system that assures good housekeeping.

If you don't, then tens or hundreds or thousands of people in your organisation are probably wasting time visiting the Internet to get the same information multiple times, saving it in multiple places, and ending up not finding it when they desperately need it — and searching the Web all over again and wasting more time.

You can save millions of dollars by using the right software system and content provision services, which will not only drastically reduce the wastage of time and free up executive time for more important work, but also give the organisation many more degrees of freedom and flexibility in utilising the information that it already has, and adds every day, and in adding value to it to serve your corporate objectives.

* Kiron Kasbekar is Managing Director of The Information Company Private Limited. A former Editor of *The Economic Times*, Bombay, Business Editor of *The Times of India* and Managing Editor of *Business India*, he is also a pioneer in India in creating business databases.

Contact:
The Information Company Private Limited
606, Aggarwal Trade Centre, Tower A
Sector 11, CBD Belapur,
Navi Mumbai - 400614, INDIA
Tel: 91-22-2756 4536 / 4537 / 4538,
Fax: 91-22-2757 1998
Website: www.informachine.com
Email: marketing@ticworks.com