

Chapter 1

Search must work

In this chapter:

- The evidence that search can have a significant impact on business performance
- A brief history of the development of search technology
- Understanding how people go about searching
- The potential for the role of information discovery manager

Searching but not finding

In July 2005 the Quarterly Survey from McKinsey Consulting¹ reported on the Global Executive Survey that the company had conducted among 7800 executives in 132 countries, a fifth of them at Chief Executive Officer (CEO) or Chief Information Officer (CIO) level. Overall, 29% of CEO/CIO-level respondents and 40% of other senior managers reported that it was difficult to find information on which to make company-wide decisions. This is a very worrying finding. Companies are flying blind, and making highly risky decisions without being able to find information that they have already created and stored.

Among many others was a survey carried out in 2004 by Vascom Bourne on behalf of Inxight Software among IT Directors in the UK financial services sector. According to the survey:

- 73% of respondents said that the main barrier knowledge workers face in sharing corporate information was not being able to use one information retrieval tool to capture data across several repositories.
- 58% of respondents also said that their company's search tools were ineffective at sourcing information quickly and efficiently.
- 66% of the companies interviewed said employees were regenerating information simply because they were unaware that the documents already existed.

For a number of years Susan Feldman and her colleagues in the Content Management and Retrieval Solutions research group of IDC (International Data Corporation)² have been surveying the time taken to undertake a range of office tasks. Their latest analysis indicates that, on average, an office worker spends 9.5 hours a week searching for information, but, of this time, 3.5 hours is wasted in not being able to find the information required. As a result, either this information has to be recreated or a decision must be made on the basis of inadequate information.

A major change over the last decade or so is that the value of unstructured, text-based information has increased substantially as organizations strive to enhance their competitive position through information and knowledge. The days when an organization depended solely on information contained in highly structured databases of client information are long gone.

The situation is getting worse by the day, a result of a continuing increase in the use of e-mail to circulate documents, the growth of content being published on an intranet through the adoption of content management software, and the need to ensure that an organization is compliant with governance regulations, such as Freedom of Information legislation and the Sarbanes-Oxley requirements. In the post-Enron business world not being able to produce documents showing that an organization has acted within relevant governance standards and guidelines

could have very serious implications for that organization. The risks arising from not being able to find information are very great indeed.

E-mails are a major source of information loss. Even if the information contained in an attached document is available from a file server, the e-mail itself usually contains a wealth of knowledge about the contents of the document, perhaps reminding readers that the table on page 6 has the wrong units on one of the axes, or highlighting a website that would provide background information on the subjects covered by the document.

All this is lost if e-mails cannot be searched. Many companies are reluctant to do this because employees also use e-mail for personal purposes, and perhaps for criticizing others in the organization. Because of the lack of a clear and enforced policy on e-mail use, the organization is at serious risk from not being able to find the information and knowledge contained in e-mail messages.

Even where there is currently some search function on a website or an intranet the organization acknowledges that it does not work and yet fails to do anything about the situation. The problem is that Google has brought search centre-screen, and provides a search benchmark that few can match or even aspire to. But users of websites and of intranets now expect there to be good search functionality, and when this is lacking begin to wonder just how committed the organization is to meeting visitor and employee expectations.

The level of investment by organizations in search technology is still very low. The current world market for information retrieval software is only around \$700 million, which in IT market terms is very small indeed. Sales of enterprise resource planning software applications are running at an annual level of \$20 billion.

What are the reasons for this state of affairs? Some clues can be found in a survey carried out in late 2005 by the Ark Group of over 500 companies.³ The chief obstacles to developing an effective search and retrieval strategy were cited as (in decreasing order of priority):

- lack of metadata and poor metadata management
- taxonomy development and maintenance
- integration with existing systems

- legacy databases and applications
- demonstrating return on investment
- teaching staff effective search techniques
- securing senior management buy-in
- geographically distributed infrastructure and content.

The issue that becomes apparent from this list is that the solutions to these problems do not lie within the ambit of any one department or Board member, so finding a search champion is very difficult. Certainly IT departments have an important role to play, but are not likely to have the detailed knowledge of the business requirements to select the optimum solution, and of course have probably never selected search software before.

Senior managers are likely to say that all they want is Google, which of course they can now have, even if the search algorithms are different from those used in the Google web search engine. Among business units the requirements of human resources (HR), sales, marketing, planning, research, customer services, and indeed any other department are not only likely to be different but to be expressed in different ways, and in the end it becomes clear that there is very little understanding of why and how people search for information.

One of the key issues with search is that people should trust it. One project we undertook for a major financial institution gained a 55% response rate from employees, and this was probably because we asked them if they trusted the search feature on the organization's intranet. The clear answer was No! If search is not trusted then it will not be used. There is no point in someone searching for information and finding that the search implementation is not giving results they can trust. In a website the visitor will move to another site, and inside an organization either the work will be redone or time will be wasted e-mailing or phoning around for the information.

The objective of this book is to ensure that benefits, risks and issues around desktop, website and enterprise search are fully appreciated.

The use of computers to search through text documents is not new, and indeed the origins of the way in which most of the current search products

work can be traced back to technology innovations on online bibliographic search services in the 1960s and 1970s. Many of the lessons learned by these search pioneers seem not to have been taken into account.

The advent of Eureka

In 1980 Sir Tim Berners-Lee had the moment of inspiration that led to his initial development of hypertext, but it was a decade later that he realized how the internet could be used to create a worldwide web of information. The first website⁴ was set up at CERN and went online on 6 August 1991. It provided information about what the world wide web was, and described the basic features of a browser and a web server. The simplicity of the concept soon led to a rapid growth of websites, and a concomitant growth in indexes to these sites to enable users to find information. However, website growth soon expanded so rapidly that the indexes could not keep up with the situation.

Boston, Massachusetts, could arguably lay claim to being the epicentre of the information revolution. Much of the early development of the internet took place in the Boston area in the early 1960s, and in 1994 Berners-Lee founded the the World Wide Web Consortium (W3C) at the Massachusetts Institute of Technology. At the same time Digital Equipment Corporation (DEC), based not many miles away from MIT, were wondering just what to do with a very high-performance chip set they had developed and codenamed Alpha.⁵ Although the performance was quite outstanding, so was the amount of heat that the chip gave off, and that was a major problem for commercial deployment. The solution came from Paul Flaherty, a research engineer working for DEC in Palo Alto, who realized that the power of the Alpha chip could be harnessed in servers that would search the web. In December 1995 DEC launched their AltaVista search engine. This used a fast, multithreaded crawler and an efficient search back-end running on the Alpha-based servers. At the time of launch it had crawled and indexed 16 million pages. Yahoo! had been launched a few months earlier, but at that time was primarily using well established concepts of categorization to provide access to the web.

Of course, using computers to search for information was nothing new. As is now recognized, much of the fundamental work was carried

out in the UK during World War 2 by the code-breaking community at Bletchley Park, and by the early 1960s the first online information retrieval systems were in the prototype stage. However, these systems were designed to be used by information professionals steeped in Boolean algebra and with the training to be able to evaluate the information that the systems presented them with. Much of the innovation that led to these services came out of System Development Corporation, in particular from Carlos Cuadra, and also from Lockheed, where Roger Summit played a key role. Both companies launched commercial services in 1972.⁶ These search services also came at a cost, because apart from connection and processing costs the publishers who owned the databases were entitled to a royalty.

With the advent of the web, where information was published free, a different approach was not only possible but essential. AltaVista changed the model and put search at the disposal of anyone with a connection to the web, who could search as many times as they wished to without charge.

For various reasons AltaVista gradually lost its market dominance, a process that was accelerated by the launch in 1998 of Google, with its innovative approach to the ranking of relevant search results. Google brought search centre-screen, and through some very sophisticated server technology and some very neat applied mathematics created a totally scalable service that now indexes billions of pages. For three decades from 1960 the top search experts worked in the IT industry with companies such as IBM and DEC. From the 1990s onward they have been working for Google, Yahoo! and Microsoft, a trio often known as GYM, and these are the companies that are setting the road-map for search technology and performance.

Of course, the work that was being undertaken at IBM and other IT companies was aimed at providing organizations with the ability to find information from the rapidly increasing collections of text documents being generated initially by word-processors and then PCs. The problems here are not on the scale of the web, but are equally challenging.

The four dimensions of search

There is a much-used matrix about knowledge that was famously used by former US Secretary of Defense Donald Rumsfeld in commenting on

intelligence failures in the Iraq conflict. Search supports four knowledge functions:

- **We know what we know.** We use search even when we know the answer to a question. That is because we like to have the assurance that we have not been overtaken by events, and we can go into a meeting confident that we have found all the information arising from a particular project or business initiative.
- **We know what we don't know.** The core function of search has always been to help us add to the knowledge that we have. We don't know the date of the launch of Alta Vista, and we need to find out for a presentation that we are giving. The challenge here is to be able to tell the search engine what we do know, so that we don't end up with a lot of irrelevant information.
- **We don't know what we know.** Our memory is far from perfect, but often we need a pointer of some sort to jog our grey cells into action: 'Of course I know that - how could I forget?'
- **We don't know what we don't know.** The biggest challenge of all for a search engine is to help in situations where we don't know what we don't know. Amazon trades on this in a very clever way by suggesting books that other readers of the book of our choice have read, as a way of informing us about things we did not know. Clustering and visualization technologies can help alert us to areas of knowledge that we did not know existed.

In probably most circumstances search cannot be accomplished by a single query: the user has to have a dialogue with the search application that can sometimes turn out to be quite a random walk towards the eventual location of the information required. Marcia Bates has described this well in her berry-picking analogy, in which the searcher obtains an initial set of documents/information, considers them, and then constructs a new search statement.⁷ Another, complementary, approach comes from Donna Maurer,⁸ in which she proposes that there are four approaches to searching, which are set out in her seminal paper:

- 1 Known-item.** Known-item information seeking is the easiest to understand. In a known-item task, the users:
 - know what they want
 - know what words to use to describe it
 - may have a fairly good understanding of where to start.
- 2 Exploratory.** In an exploratory task, people have some idea of what they need to know. However, they may or may not know how to articulate it, and, if they can, they may not yet know the right words to use. They may not know where to start to look. They will usually recognize when they have found the right answer, but may not know whether they have found enough information.

In this mode, the information need will almost certainly change as they discover information and learn, and the gap between their current knowledge and their target knowledge narrows.
- 3 Don't know what you need to know.** The key concept behind this mode is that people often don't know exactly what they need to know. They may think they need one thing but they actually need another; or they may be looking at a website without a specific goal in mind.
- 4 Re-finding.** This mode is relatively straightforward: people looking for things they have already seen. They may remember exactly where it is, remember what site it was on, or have little idea about where it was.

The alignment of this approach with the matrix above is very close. Another important study on the way in which people search for information has been the work on information scent by Chi and his colleagues at the Xerox Palo Alto Research Center (PARC)^{9,10}

In summary, designing effective search systems involves an excellent understanding of the way in which our brains process information, and in particular the limitations of short term memory. This issue of short term memory has a major impact on the design of search results pages.

Search is quite a complicated technology and, as is described in more detail in Chapter 2, comprises at a minimum:

- indexing
- query management
- ranking of results
- results formatting.

Every search vendor has their own views on the most effective way to undertake these four operations, which are all interlinked. Only through careful testing against sample sets of documents and queries, and then by constant tuning of the search engine, can the full effectiveness of the search engine be obtained. There is a need to support a dialogue with the search user, and usability has to come at the very top of the implementation process. The failure of many search implementations is that they assume that all users search the same way. The truth is that every user is different, as we are always looking to add to what we already know, and only we know what we know, provided we can remember what we know.

Findability

Over the last four decades a considerable amount of research has been carried out into all aspects of information retrieval, as even a cursory glance at journals such as the *Journal of Information Science* or a look through the papers at the annual Text Retrieval Conferences (TREC) will show.¹¹ There is a danger of taking search out of context. All too often the justification for implementing a search solution for a corporate website or an intranet is that people cannot find the information they want through the site navigation. Once search has been implemented, the alarming discovery is then made that there are some basic underlying issues of poor content management and metadata management, and all that the search engine has done is bring a magnifying glass to these issues.

Search must be part of a total information discovery strategy, and someone who has thought a lot about this issue is Peter Morville,¹² with his concept of 'findability'. This he defines as:

- the quality of being locatable or navigable
- the degree to which a particular object is easy to discover or locate

- the degree to which a system or environment supports navigation and retrieval.

In the end the requirement is to be able to find an object, be it a document, a piece of data or a video file. There are only three information discovery routes:

- lists, indexes and classifications
- hyperlinking to related information
- searching through a computer-created index of all relevant metadata.

The challenge for any organization is to achieve the correct balance of these routes. A website represents a known collection of objects, and so much can be done through lists and indexes, and of course through hyperlinking. Search can be important, and is used either at the outset to reach the relevant area of a site quickly, or as a fallback when all else has failed.

When it comes to internal websites - intranets - hyperlinking is much less effective because of the heterogeneous nature of the content, a higher proportion of longer documents in formats other than HTML, and a distributed content authoring environment in which no single person has a sufficiently comprehensive knowledge of the collection to be able to decide where hyperlinks should be added. In intranets search is not a nice-to-have but a need-to-have.

Search and navigation

Search is not a solution to poor information architecture. Effective search has to be integrated within the overall information architecture of a website or an intranet, so that there is a seamless path between the structured navigation, hyperlinks and the search function. In presenting the results of a search it can be of value to include the URL of the displayed item in a way that enables the searcher to realize that there is a section of the site that they have not visited. This requires the URLs to be short, structured and intuitive, features that are often missing with the increased use of dynamic page publishing and portal applications where

the URL also contains session information that is of no assistance with site discovery.

The implementation of a new search engine may require a redesign of the site, or at least sections of it, and this can add to the cost and implementation schedule of the release of the search function. However, just adding a new search box is not going to obtain the best return on the investment.

The role of the Information Discovery Manager

The information profession seems to be constantly trying to define a role for itself in 21st century organizations. Certainly many intranets are managed by information professionals, and they are also involved in websites. To do this they need new skills in websites technology, but they are missing an obvious opportunity, that is, to support the development of search within the organization - not just for the intranet, but across all applications and requirements. They already know how search should work, and they have the experience gained in working with online database services. They are heavily involved in web research, and understand how taxonomies and classifications can be used to enhance the information discovery process.

Indeed, the opportunity is there for a new position inside an organization, that of Information Discovery Manager. The scope is enormous and business-critical. The roles that are well suited to an information professional acting as Information Discovery Manager include:

- developing and maintaining taxonomies
- developing and supervising usability tests
- developing metadata schemas
- ensuring that external information resources are integrated into the search experience
- identifying 'best bets' documents for important searches
- identifying the scope of test collections of documents for use in the evaluation of search products
- managing the search 'helpdesk'

- monitoring developments in search technology and the search business
- reviewing search logs to develop search enhancements
- teaching staff effective search techniques.

The remainder are clearly the responsibility of the IT department, and with search in particular there is an important role for IT to provide an adequate technology infrastructure to support search applications.

References

- 1** www.mckinseyquarterly.com/home.aspx.
- 2** www.idc.com.
- 3** Ark Group (2005) *The Age of Search: intelligent retrieval and analysis*, London, Ark Group, www.ark-group.com.
- 4** <http://info.cern.ch/>.
- 5** www.washingtontechnology.com/news/10_23/news/10125-1.html.
- 6** Bourne, C. P. and Hahn, T. B. (2002) *A History of Online Information Services*, Cambridge MA, MIT Press.
- 7** www.gseis.ucla.edu/faculty/bates/berrypicking.html.
- 8** www.boxesandarrows.com/view/four_modes_of_seeking_information_and_how_to_design_for_them.
- 9** www2.parc.com/istl/groups/uir/publications/items/UIR-2001-07-Chi-CHI2001_InfoScentModel.pdf.
- 10** www.steptwo.com.au/papers/kmc_informationscent.index.html.
- 11** <http://trec.nist.gov/>.
- 12** Morville, P. (2005) *Ambient Findability*, Sebastopol CA, O'Reilly Publishing, www.oreilly.com.