Advances in search engine technology and their impact on libraries

Dosapaty.Vasu
Librarian,
TKR Institute of Management & Science, Hyderabad.

Abstract:

Libraries see themselves as central information providers for their clientele, at universities or research institutions with the development of the World Wide Web, the "information search" has grown to be a significant business sector of a global, competitive and commercial market. Powerful players have entered this market, such as commercial internet search engines, information portals, multinational publishers and online content integrators. If libraries do not want to become marginalized in a key area of their traditional services, they need to acknowledge the challenges that come with the globalization of scholarly information, the existence and further growth of the academic internet.
Keywords: library, information search, globalization, information portals

INTRODUCTION:

With today's instant anywhere-anytime access to Google, Bing and Wolfram Alpha, where searching for information takes a few micro seconds via an internet-connected device, some people regard physical libraries as a quaint relics of a forgotten age. Looking at the practice of today's digital library portals we get the impression that the internet is almost non-existent in the academic resource discovery environment. What we find are online library catalogues, electronic journals and (sometimes) e-books, which are mainly digitally converted print materials that have traditionally been the focus of library acquisition policies. Also databases have been well known for a long time. Content is generally delivered through well-established service channels by publishers, book-houses or subscription agencies.

The digitization of publishing and the advent of the World Wide Web have resulted in the proliferation of a vast amount of content types and formats that include, but are not limited to, digitized collections, faculty and research groups' websites, conference web servers, preprint/e-print servers and, increasingly, institutional repositories and archives, as well as a wide range of learning objects and courses.

If these resources are registered by a library at all, then they are in the form of separate lists of links or databases, but are not integrated into local digital library portals.

Literature review:

As a survey carried out at Bielefeld University in November 2002 revealed, students still make intensive use of the online library catalogue, but they would much prefer to access the catalogue through a "Google-like"-interface. The simple reason why they still use the online catalogue is that, for this information type, they don't have an available alternative, as internet search engines usually don't cover the so called "deep" or "invisible" web. In any area where students think that they can find information, especially when they are looking for documents and full text, general search engines are even now much more popular than databases that have been made available through libraries. And it is only because of their level of experience and a certain habit that researchers still use databases, e-journals etc. that are not indexed by internet indexes. But a new generation of researchers will be coming in a few years time that will have grown up with popular internet search tools.

Just as users like the ease of phrasing and submitting a search query, they also like the flexible and responsive display of result sets. Superior performance and the size of internet search indexes are most impressive to them. Already published in 2001, a white paper from Michael Bergman on the "Deep Web", highlights the dimensions we have to consider. Bergman talks about one billion individual documents in the "visible" and nearly 550 billion documents on 200,000 web sites in the "deep" web. The exponential growth since 2001 can be read from the fact that in May 2004 Google gives the size of their index (i.e. visible web content) with more than 4,2 billion web pages (compared to 3,3 billion web pages in 2003). In May 2004, 167 million science-specific Web pages have been indexed by Scirus, that are roughly 4% of the public Google-index (4.2 billion web pages). Although Scirus includes some "invisible" resources, the majority of the information has

been crawled on web sites that are marked as "scientific" through their domain names. Taking aside all the vagueness of these estimations one might apply the 4%-factor on the size of the invisible web in 2000 (i.e. 550 billion web pages) and receives the impressive figure of 22 billion web pages that include scientific content.

The impact of internet search engines on libraries:

It is a fact that with the advent of the World Wide Web, the information "search" has grown to be a significant business sector of a global, competitive and commercial market. Libraries are only one player within this market. Other stakeholders include, but are not limited to, publishers, online content integrators and commercial internet search engines ("information.coms").

In any market situation it is of very importance to take a close look at potential customers and their usage behavior. For librarians this might sound obvious as it is their genuine perception that they consider implicitly the demands of users—or rather what they consider to be the demands of their users. But the new, competitive situation forces libraries to see things much more from the perspective of the user. First of all, this is an acknowledgement that, particularly at universities, libraries deals with a range of users with often different usage behaviors'.

It almost goes without saying that an undergraduate has other demands for information than a qualified researcher, and their usage behaviors can vary substantially. Young undergraduates will try much harder to transfer their general information seeking behavior (using internet search engines) to the specific, academic environment, while established researchers have better accommodated the use of specific search tools. Before the WWW had been developed, this differentiation was, from the librarian's point of view, only relevant with respect to the level of training that various user groups required in order to use the library's resource discovery tools (printed catalogue, online catalogue, digital library portal).

Today, with a whole range of general search engines available, users have the opportunity to use other catalogues (public or academic), and portals than those found in the library. Library users have been "empowered" by Google-like search engines to make their own choice about a search tool and to approach the world of information without any training. While librarians are mainly worried about the quality of information resources that are covered by mainstream search indexes, their users love these new tools and they would like to use them for any type of information search.

Hesitation from libraries:

It is possible to identify reasons why libraries are hesitant to take action to change this situation. As a matter of principle libraries rank locally held collections and resources much more highly than remote resources, as the size of local collections has always been one indicator of the importance of a library. Libraries still see themselves as a place of collections rather than as an information "gateway". Other concerns of libraries are grounded in the fact that there is no guarantee that a remote host will maintain its resources in the long-term. Thus gateways to remote resources always have to face the potential problem of dead links. However, long-term accessibility is one of the basic values of libraries, and procedures need to be set up that will ensure that even remote repositories can be accessed in the long-term.

Other reasons may include a natural resistance to the change of established acquisition procedures and workflows, as well as the complex combination of skills and competences that are required for "acquiring" remote resources, such as subject expertise, technical knowledge and traditional acquisition skills. This new type of "acquisition", introduced as a regular workflow within a library, would require some re-organization of current structures with potential implications for costs and resources.

Are libraries aware of the incredible volume of academic content that is available on the web?

While libraries concentrate on the building of local digital library portals and simultaneous searches across a selected number of licensed and free databases, do they see the incredible volume of academic content that is already available on the web? Although there are no reliable figures on the overall volume of web content there have been some studies that give estimations. For the research and teaching community the "invisible" web is of specific interest as it includes to a major proportion (high) quality content in free or licensed databases, primary data (e.g. meteorological, financial statistics, source data for bioresearch and so forth) or the huge and still increasing range of cultural and historical resources that are being digitized. There are, again, no reliable figures on the actual size of the academic web but the size of Scirus , the "science-specific search engine" of Elsevier might serve as a pure indicator for the volume dimension.

The vision...

Instead of a highly fragmented landscape that forces users to visit multiple, distributed servers, libraries will provide a search index, which forms a virtual resource of unprecedented comprehensiveness to any type and format of academically relevant content. Libraries liaising with other partners are contributing ultimately

to an open, federated search index network that will offer an alternative to the monolithic structures of current commercial information.com indexes.

This unique resource will not form a minor segment within a commercial internet index, which lives from and is often heavily influenced by the advertisement industry, with their very specific rules about relevance and sustainability of information. Libraries will offer a long-term, reliable search service, which comprises high-quality content for the research and teaching communities.

Libraries are increasingly hesitant to support big, monolithic and centralized portal solutions equipped with an all-inclusive search interface which would only add another link to the local, customer-oriented information services. Future search services should be based on a collaboratively constructed, major shared data resource, but must come with a whole range of customizable search and browsing interfaces that can be seamlessly integrated into any local information portal, subject specific gateway or personal research and learning environment.. Libraries using the new search index must be able to select only those data segments that are of relevance specifically to their local or subject specific clientele, and search and browsing interfaces need to be customizable according to the local "look and feel" or discipline specific navigation mechanisms.

The new, academic search index should come with the ease of handling and the robustness and performance of Google-like services but with the relevance and proven ("certified") quality of content as it is traditionally made available through libraries. While undoubtedly successful in offering integrated access points, from the library point of view one gets the impression that there is still some development to be done in order to build real end-user services that find the full acceptance of researchers and students.

In the era of popular internet full text search indexes these projects are focusing mainly on metadata by giving reference information about the resource (e.g. a certain server or database) rather than searching within the content sources (such as the full text itself). The records of all these portal databases, which usually describe intellectually selected content sources, can of course be used as a valuable starting point for the proposed discovery of the academic web.

Where internet addresses are included in these records they can serve as starting URLs for web crawlers and other data aggregation tools that come with search engine technology. It should be noted however that the major work for libraries building an academic web index will begin after the resource has been located, as a major proportion of the content in the academic web can not be aggregated by standard crawling mechanisms.

This is why this part of the internet is called the "deep" or "invisible" web, and it comes, in particular in the academic environment, with an almost endless variety of data formats and technological implementations of databases and content servers.

Conclusion:

This paper advocates a concerted initiative of the library community to pick up state-of-the-art search technology and build reliable, high quality search services for the research and teaching community. This effort is not intended as competition to other commercial services, but it represents a natural continuation of traditional library services in a globalised academic information environment.

An academic internet index network, driven by the library community as sketched out before, can best meet the specific requirements of the scholarly clientele by providing comfortable, reliable and integrated access to high quality content.

But in order to realize this new service libraries are required to look beyond their current information infrastructure to learn from mainstream internet index providers who have become so popular through innovative technology and a dedicated end-user driven approach.

How can libraries proceed:

The library community needs to acknowledge the relevance of a new action plan in order to improve current search services. The impression is that many libraries "somehow" see the need but it's still unclear for them how to address the problem. Current pragmatic approaches to make academic content available to commercial internet indexes should be seen only as a first step on the way to a new service that is driven by the libraries themselves.

It is expected that other libraries will start to create their own local search engine infrastructures in order to build further indexes. An (informal) network or forum will be formed where knowledge, content and tools can be shared.

The need for trans-national action has never been so obvious as it is now, and it is hoped that funding organizations in many countries will acknowledge the dimension of this undertaking and give the support libraries need to fulfill their mission: to discover the rich wealth of the academic internet for the benefit of the international research and teaching community.

References:

1.www.campbellcollaboration.org/lib/download/969/.

2.www.dlib.org/dlib/june04/lossau/06lossau.html.

3.www.researchgate.net/.../Impact_of_search_engines_in_Library_and_Inf...

4.www.canis.illinois.edu/news/Computerintro.pdf.

5.www.webpages.uidaho.edu/~mbolin/krubu-osawaru.html.