Google Scholar versus Metasearch Systems

Tamar Sadeh, Ex Libris

Workshop at the 29th ELAG conference, May 31-June 4, 2005

Background

No doubt that Google as a Web search engine has had a great impact on all those who search for information on the Web. Its instant response, huge repositories, sophisticated search mechanism, and relevance ranking have combined to make it the number one Web search engine.

Over the last year, Google launched several exciting products; one of them is Google Scholar. Aiming to provide a single repository for scholarly information, Google Scholar enables users to search for peer-reviewed papers, theses, books, preprints, abstracts, and technical reports in many academic areas. Furthermore, according to information released by Google, Google Scholar arranges results by relevance, taking into account the full text of each article as well as the article's author, the publication in which it appears, and the number of times that it has been cited in scholarly literature. Equipped with this unique ranking process, unparalleled hardware resources, sophisticated crawling techniques, and well-established collaboration with publishers, Google is positioning Scholar to be an essential resource for the scholarly environment. In the not-too-distant future, Google is likely to be facing rivals such as MSN and Yahoo, who will probably offer similar products.

In light of these recent developments, we need to examine the historical roots of this type of searching.

First, let’s clarify the difference between a metasearch system and a federated search system, such as Google, and make sure that we share the same terminology.

Metasearching, also known as integrated searching, simultaneous searching, cross-database searching, parallel searching, and broadcast searching, refers to a process in which a user submits a query simultaneously to numerous information resources. The resources can be heterogeneous in many respects: their location, the format of the information that they offer, the technologies on which they draw, the types of materials that they contain, and more. The user’s query is broadcast to each resource, and results are returned to the user.

The development of software products that offer such metasearching relies on the fact that each information resource has its own search engine. The metasearch system transmits a user's query to that search engine and directs it to perform the actual search. Upon receiving the results of the search, the metasearch system displays them to the user. This process involves the adaptation of the query to the specific requirements of the search engine at the target’s end, as well as the conversion of the results to a unified format. The unified format later enables the metasearch system to process the results further—including displaying them in a consistent manner and merging and de-duplicating them.

We can describe metasearching as "just-in-time" processing. That is, instead of preprocessing the data, the metasearch system processes it only when the user launches a query. Metasearch systems, therefore, hold information about how a resource can be searched and how results can be extracted from it, but they do not contain any of the data that is stored in any of the resources they can access.

In federated searching, a wealth of information is incorporated into a single repository that can be searched. In this scenario, the information is processed prior to the user's search. From the end user's point of view, the two methods may seem similar, but in fact they differ in many respects. Such preprocessing, which we can describe as "just-in-case" processing, opens new horizons regarding the search techniques and the presentation of the results. We will discuss these new capabilities and the differences between the two types of technologies during the workshop sessions.

Looking back a few years, we can clearly see that the need for a single search interface to multiple resources is not new, and, in fact, metasearching and federated searching have been available for quite some time. Such systems originated in various environments; for example, publishers offering numerous journals created a federated search mechanism enabling their users to search all their e-journals. A good example is Elsevier's ScienceDirect. As Elsevier acquired other publishers, it was able to add their journals to the same platform.

Database vendors developed similar mechanisms. For example, SilverPlatter provided a single interface to more than 300 databases that they published (note that in this case the various databases were not merged into one, yet the interface could search them all). Commercial organizations were not the only ones that felt the need for federated searching; other types of institutions created a local environment based on federation. For example, the Los Alamos National Laboratories and OhioLink in the United States, the University of Toronto in Canada, the Technical Knowledge Center and Library of Denmark (DTV), and the Max Planck Society in Germany all offer large, diverse collections of e-journals that they store locally. These institutions have implemented federated searching to provide a single search interface across their electronic collections. However, not all organizations have the resources to adopt this just-in-case approach.

Federated searching has not adequately addressed the needs of researchers. The number of heterogeneous resources that institutions offer their users has increased rapidly, and a single federated searching system can serve only as a partial solution. Further step toward metasearching was taken by library system vendors, who used Z39.50 as both a gateway and a server and thus provided access to library catalogs. But, once again, this solution could not scale up. Hence, we saw the emergence of dedicated metasearch systems as we know them today.

The quick acceptance of metasearch systems by the market indicates that libraries, indeed, have a need that these systems can fulfill. For example, nearly 500 institutions have acquired the Ex Libris MetaLib® system since 2001, and many other such metasearch systems are offered in the marketplace. The ability to provide a single, friendly interface to multiple resources enables libraries to better address the changing expectations of their users, users who in the meantime have become accustomed to Google and Amazon.

Libraries have not only adopted metasearch systems at a rapid pace, but they have also advocated the development of new standards related to the metasearch process and are sharing their concern with information providers and metasearch system vendors. This concern led to the formation of the NISO Metasearch Initiative, whose aim has been to provide the industry with a set of standards that will facilitate and optimize metasearching. This NISO initiative has been the focus of much discussion in the last couple of years, and apparently numerous stakeholders—publishers, librarians, and metasearch system vendors—see the significance of forming standards in this area.

Of particular interest are the Semantic Web developments spearheaded by Tim Berners-Lee and the Worldwide Web Consortium (W3C). The Semantic Web approach, if it materializes, can facilitate the interaction between a metasearch system and any number of target resources without requiring prior programming for each target resource. The ideal solution is to receive resource-specific information at the time of the actual interaction—as envisioned by the promoters of the Z39.50 Explain facility—and formulate the flow of the interaction on the basis of this information.

The launch of Google Scholar is generating much interest in the industry and quite a few concerns. The main question is whether systems such as Google Scholar will completely replace metasearch systems and make them obsolete or whether the two types of systems will exist side by side.

Workshop Topics

At the workshop we will discuss the following topics:

- The concept of the infrastructure on which metasearch systems are based, as opposed to the concept of the federated search system infrastructure

- Metasearch systems: advantages and disadvantages, including the NISO Metasearch Initiative's anticipated impact on the industry once standards have been set. For example, how easily will libraries be able to discover and exchange collection-description and service-access description metadata, which is potentially also available for end users?

- Federated search systems: advantages and disadvantages of the approach of Google Scholar. This topic includes disadvantages inherent in the nature of the system as opposed to those related to Google Scholar's implementation of the concept.

- Concerns of libraries regarding Google Scholar, such as:
  - The scope, coverage, and accuracy of the content
  - The completeness of the data
  - The focus on specific disciplines
  - Selective searches
  - Relevance ranking
  - Integration with the local environment, such as with local authentication systems
  - Local branding
  - Integration with link servers
  - The incorporation of holdings information in Google Scholar

- ❖ Output tools, such as RefWorks, EndNote®, and ProCite®
- ❖ Business models, such as the use of advertisements
- ❖ Lack of control over all the above—perhaps the most important concern of all

- Looking ahead: What tools do end users need to conduct their research? What role do librarians play in the research process? Does Google Scholar provide an environment that librarians approve of? If not, will they come to approve of it in the future? Will end users search only via Google Scholar—or similar systems—or will they also still need the types of systems that libraries provide and that use available metasearch tools?

Suggested Reading

1. Google Scholar: A review by Peter Jacso. Péter's Digital Reference Shelf, December 2004. http://www.gale.com/servlet/HTMLFileServlet?imprint=9999&region=7&fileName=/reference/archive/200412/googlescholar.html

2. Is MetaSearch Dead? A presentation by Roy Tennant of the California Digital Library, March 2005. http://escholarship.cdlib.org/rtennant/presentations/2005niso/