



IS414: Search Engine Technologies

**Searching across academic databases:
A comparison between three approaches**

NG Woon Bock
YU Shuli

Introduction

The library at Singapore Management University (SMU) subscribes to a number of academic databases which contain collections of journal articles, conference proceedings and working papers among other documents. These are frequently accessed by students and faculty who require the documents in their course of their daily research and work.

Currently, the process of searching and retrieving articles is tedious as it is not possible to query different databases at the same time. For example, documents pertinent to the topic of “Search Engine Indexing Technologies” are present in numerous journals by different publishers who each maintain their own database, so these relevant articles could be spread across various databases like JSTOR, ScienceDirect and EBSCOHost. In order for a user to obtain articles from a variety of sources, he has to access each database individually from the SMU library website and then perform his search repeatedly, across every database that might contain relevant documents.

This method is laborious and time-consuming, and often results in users restricting their search to a small subset of the available databases, as it would be too troublesome to repeat the search for all the known databases. As such, there is a high possibility that documents relevant to a user’s research are neglected because they are found in some of the smaller or lesser known databases.

Hence the need for an integrated searching experience. After speaking to the librarians at SMU about this problem, we understand that they are looking to alternatives to the current search process, and have short listed a number of approaches that could address the above problem. This document aims to describe and evaluate these approaches in order to determine the one that can best address the needs of SMU.

The Different Approaches

Three main approaches: 1) In-house search and indexing; 2) Library Links Program (using Google Scholar); and 3) Metasearch have been identified as viable alternatives to address the current disparate searching process.

1) In-house search and indexing

This approach involves indexing and storing the documents in-house from all the databases that SMU Library subscribes, and the providing a search application to search through these indexes. (See Fig 1 for the logical information flow.)

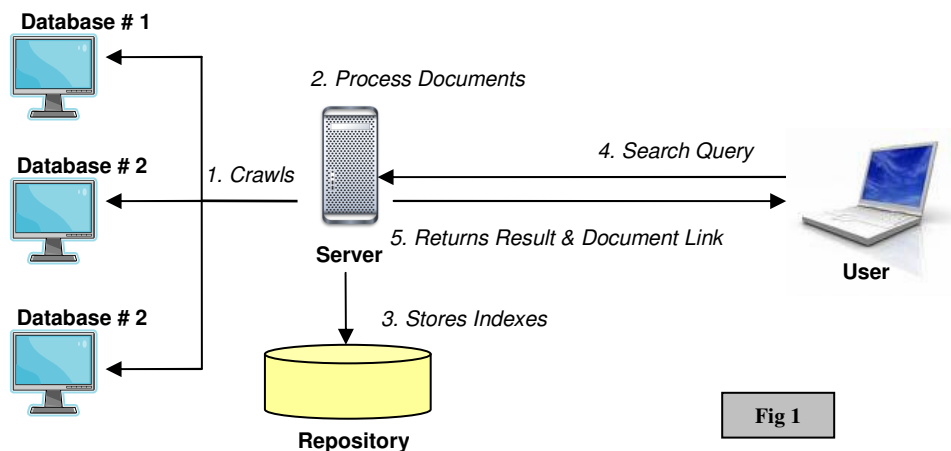
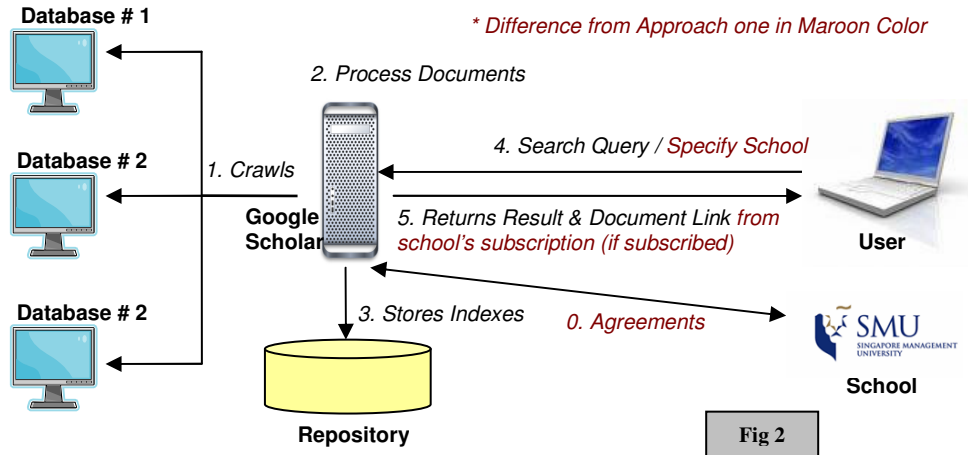


Fig 1

In this process, SMU would possess great amount of control, and can use either commercial or open source search engine applications and customize to their needs. Documents from all databases would be stored within SMU’s server for indexing purpose, and the average user would be able to search for his requests through SMU’s point of entry.

2) Library Links Program (using Google Scholar)

This approach leverages on readily available services on the Internet, and involves forming an agreement with Google. The innate process works similarly to the previous approach, except that the user now queries Google Scholar instead of the school's search engine. (See Fig. 2) The user could either use Google Scholar from within the school campus, or specify in the preferences that he is from the specific school (in our case, SMU).



The agreements may include defining the IP address of the school, the subscribed databases, and the customized links to log into the subscribed databases which would differ for different school. The fact that Google Scholar possesses enormous bargaining power and value to its partners is a key reason why many database owners enter into these agreements which allow Google to index their documents.

Results returned thus would return the links in the manner depicted in Fig 2.1, where "Find it @ Singapore Management University" indicates that the article is being subscribed to by the school.

[book Agriculture and Income Distribution in Rural Vietnam Under Economic Reforms: A Tale of Two Regions - all 19 versions »](#)

D Benjamin, L Brandt, William Davidson Institute - 2002 - bus.umich.edu

... in Rural Vietnam under Economic Reforms: ... Page 3. Agriculture and Income Distribution

in Rural Vietnam under Economic Reforms: A Tale of Two Regions ...

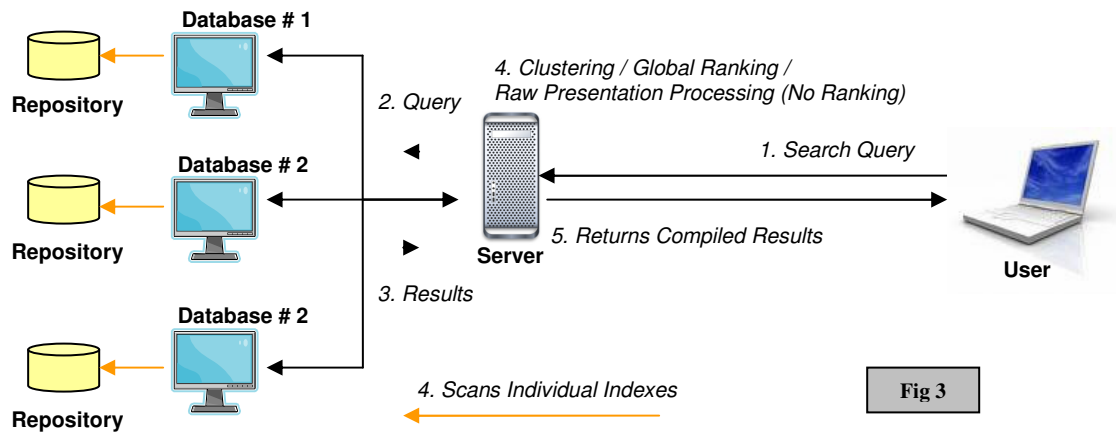
[Cited by 42](#) - [Related Articles](#) - [View as HTML](#) - [Web Search](#) - [Find it @ Singapore Management University - Library Search](#)

Fig 2.1

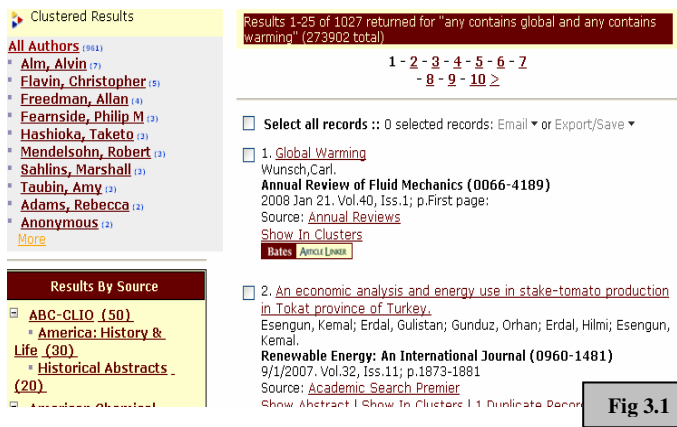
* note, this illustration was manipulated. SMU has not liaised with Google Scholar as of today

3) Meta-Search Approach

This approach is often used interchangeably with the term 'Federated Search'; and for the purpose of this discussion, Meta-Search and Federated search will be treated as similar. Here, a user first queries a single entry point in the school, and the school's sever would in turn query the database's search engines and returns the results from each one. The results are then compiled, either into multiple lists or a single ranked list; whichever the method, this approach allows users to search for contents in multiple databases at one time with only 1 click. As explained in Fig. 3, the onus is on each database to return search results, and the school's server would be the intermediary between the 2 entities.



This approach has been used in several institutions in Singapore such as Temasek Polytechnic and National University of Singapore, and there are a wide range of commercial vendors and solutions available, such as Ex Libris Metalibⁱⁱ, Innovative Research Proⁱⁱⁱ and AGen Search^{iv}



This approach's other key advantage is in reaching the "deep web" – content that are not indexed in common Internet Search Engine, as this remains Google Scholar's key agenda.

In Fig. 3.1, Serials Solution provides a form of meta-search approach which allows clustering by Source, Author, and global ranking to the documents. For the purpose of this discussion, repositories for indexing should not exist within the school's server (otherwise, it would be akin to in-house search and indexing, the first approach).

Evaluation of the Approaches

In order to determine which approach would be suited for SMU, we have evaluated the approaches based on two main criteria: 1) Usability and 2) Implementation. Usability will refer to the ease of the user experience in using each approach, while implementation would focus on the feasibility, effort and costs of developing and maintain each solution.

The following table is a summary of the comparison between the three approaches, where 1 star represents the solution being the least effective and three stars represents the solution as highly effective for each of the criteria.

| | Usability | | | Implementation | | |
|----------------|----------------|----------------|--------------------|------------------------|---------------|-----------------------|
| | Global Ranking | User Interface | Speed of Retrieval | Licensing and Indexing | Local Control | Cost, Effort and Time |
| In-house | ★★★ | ★★ | ★★★★ | ★ | ★★★★ | ★ |
| Google Scholar | ★★★ | ★★ | ★★★★ | ★★★★ | ★ | ★★★★ |
| Meta-search | ★★ | ★★★★ | ★★ | ★★★★ | ★★ | ★★ |

Usability

Successful search applications get it 'right' for the end users by having them in mind. The effect of having relevant results right in the first page, if not the first few document results, have far-fetched benefits and is of utmost importance. Similarly, if an application takes 5 minutes to return results, or requires 3 clicks and repeatedly entering login details to access any database after receiving the results, the application would defeat the very purpose of implementation and not prove to be useful at all.

Accordingly, this section attempts to evaluate the degree of ease and satisfaction in using the proposed solutions in the perspective of an end user.

Global Ranking

An in-house search and indexing engine would fare well in returning relevant and precise results, and produce one result set only. The school's administrators would therefore be able to customize ranking and search algorithms that best suit the local context (in this case, SMU is essentially a business school) that can be reviewed frequently in accordance to analyzing search queries and behavior.

Likewise, Google Scholar will return one global set of results, which is based on several factors like the number of citations^v, as well as the article's author, the publication in which the article appeared and how often it has been cited in scholarly literature"^{vi}. Through leveraging on Google's historical search innovation prowess and capacity, it is reasonable to assume that their relevance ranking can improve leaps and bounds. Currently, although they return decently ranked results, there are minor flaws with their ranking algorithms, such as the non-existent verification of duplications of journals whose citations count towards any article. The APIs are also not released for organizations to fine tune the ranking. Nonetheless, Google Scholar remains well on course for quality results in future, especially since they are able to index a huge volume of databases.

Clearly, in the perspective of a meta-search application, ranking of results within each database is within the control of the individual databases, thereby posing a weakness. This means that the numerous result sets will be returned initially for the meta search engine to combine into one cohesive result set. However, it is known that most of the commonly used meta-search algorithms are based on on ad hoc techniques which are rather simplistic; for example, simply interleaving the ranked lists returned by the underlying search engines or ordering documents by their rank-sum over the underlying systems.^{vii} Thus, this problem could result in a global result set that lacks accuracy, or if global ranking is ignored in favor of multiple result sets or just clustered results, the eventual results returned may prove messy and overwhelming to the user.

User Interface

Completely customizable, the in-house search and indexing approach would tip the scale in this criteria scoring. Librarians would be able to fit the search into current systems, and develop interfaces that are familiar to users in the school. Since the documents may reside in the school's system, users may even bypass the formalities of entering the database system and access the documents directly. If that violates copyright issues, the school can at least facilitate the user to access the database with as little clicks as possible, through middleware. For e.g. clicking on a link in the results will enable the user to log into the database system directly. Results can also be clustered, if the school so wishes to develop the functionalities.

Meta-search applications would also provide the bulk of the benefits an in-house search and indexing application would offer in terms of user interface, except minute customization restrictions such as layouts. Otherwise, these applications are developed by vendors who are familiar with customer requirements and needs, and can be trusted upon to deliver interfaces which do not require intensive user training.

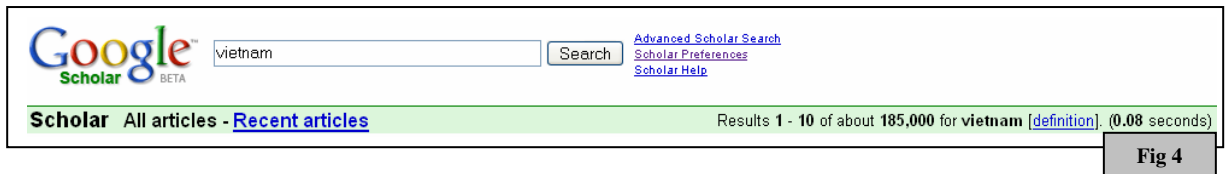
Google Scholar almost is not coupled with any above mentioned privileges. Though liaisons with the school^{viii} may offset some inconveniences for the user by creating a link to the school's database directly in results, users would still have to access Google Scholar in the Internet, since there is no

integration with the school's virtual environment (This also result in a loss of university branding). However, it is noteworthy that Google Scholar provides the same intuitive interface in the well recognized Google search which is simple in terms of design and functionality, and encourages a natural transition.

Speed of Retrieval

Assuming the user started the search within a school's network, the in-house approach would be fast, since the server would work on previously processed indexes and there is no requirement to access the Internet. The library too has the option to scale up in terms of hardware and bandwidth when the need arises, both for handling search queries, and for crawling and indexing the database's documents.

A primitive test to assess the speed of Google Scholar was conducted. See Fig 4.



A turnaround time of 0.08 seconds can hardly be considered as average for a search application results delivery. However, note that this test was not conducted fairly at a peak period where thousands of users are concurrently using Google Scholar and a peak period where Internet bandwidth would be diminishing. Results would differ in when bandwidths are highly utilized, though the difference may be negligible.

The work flow of the last approach, meta-search, would understandably means that it would be the slowest of the 3 evaluated approaches. The bottleneck in this case would be the slowest database to return a result after the school's server queried multiple database search engine (which in turn queries their own indexes) using Internet bandwidth. Likewise, because it is a "just in time" search aggregation service, time is required for the server to compile the results in the desired manner the user wish for, or according to customization instructions the developers or administrators have configured.

Implementation

The feasibility of implementing each solution is integral to the eventual selection of an approach, as SMU must have the capabilities to develop or customize the solution to interface with the external databases and its own library system, as well as the resources to maintain and update the system after it has been adopted.

Licensing and Indexing

Search engines need to have access to the pool of relevant documents, in order to generate an index which is used in the search process. At minimum, the solution would require the access and indexing of all the articles in the databases that have been subscribed. In order to do so, licensing issues must be overcome because the databases are proprietary and most of the documents under them are restricted content.

This would be the most challenging for the in-house approach, because both licensing and indexing must be initiated by SMU, as development will be done by the school. Agreements must be made individually with each of the databases that have been subscribed, so that we have access to crawl through the database and index the documents. There is also a need to determine and synchronize the method of crawling and sifting through the metadata embedded for each document, as each database would have a different format. Thus, the approach would require a fair amount of time for development due to the need to liaise with the database publishers. Also, any refusal to allow crawling and indexing would jeopardize the approach greatly, as it would decrease the recall of the

search results since there could be a number of relevant documents that were not even indexed, and thus not part of the search results.

In contrast, using Google Scholar would place the onus of liaising with the database owners on the respective vendors. Google already has entered into “numerous individual agreements with publishers to index full-text content not otherwise accessible via the open Web”^{ix}, which includes indexing academic documents from the databases required. This is possible because of the size and reach of Google – it is practical for private databases for their documents to be reflected in the search results of Google because it would bring in a greater amount of traffic. The databases can then determine if the users have the rights to view the article, and if not, they could limit viewing by directing the user to the document abstract, to entice the user to purchase the full-text version instead.

Similarly, there would not be any problem with indexing issues with the meta-search approach, because commercial meta-search solutions focus on providing an interface to search more than one database at the same time, while leaving the database to determine the method of ranking the results. This means that any form of indexing or algorithms required is hosted at the database level and does not need to be performed directly by the application.

Local Control and Availability of Source Code

Local control is important because there is a need to restrict the search results to only those subscribed by SMU library, and there might also be a possibility where the ranking algorithm should be customized to suit the main users of the application – the students and faculty of SMU. This is influenced by how easy it is to obtain and alter the source code of the solution.

Local control is not an issue for the in-house approach, as development can be highly controlled. The solution can either be based on open-source APIs and developed completely in-house, or it can be customized from the purchase of an existing search engine. Both methods would give the developers control over the source code and allow for later updates. These customizations to the ranking procedure can be built into the indexing and ultimate ranking algorithm; and any changes to the relevance rankings by the database would not affect the application because it would be relying on self-produced indexes, not that of the database or any other commercial application.

Using Google Scholar, however, would mean relinquishing control over any ranking algorithm or search restrictions. As this approach merely consists of liaising with Google to add on the “Full Text @ SMU” link for the results that have subscriptions, the eventual application used will still be hosted by Google, so no special customizations can be made to the ranking process. This is because the source code to Google Scholar remains proprietary, and Google has not released its APIs for public use. Also, the pool of documents in the search cannot be restricted to the databases subscribed, as Google Scholar will return results from its entire content pool. Even so, it must be noted that the Google Scholar does have a powerful ranking algorithm, and this approach could return additional relevant results where full-text articles can be found on the author’s or university websites, bypassing other databases not subscribed by SMU.

It would be difficult to manipulate the relevance rankings within each of the databases when the meta-search approach is used. As this method relies on using the search engines within each database, the ranking method within the individual databases cannot be altered. However, it is possible to tailor the eventual presentation of the results once they have been obtained separately from the databases. Currently, there is a wide variety of commercial meta-search products available, and most of these products allow customization after purchasing. This allows for the possibility of having features that allow result clustering by database, author, or date, or even result combination – the integration of separate results into one result set using another ranking algorithm developed in-house, or by the commercial product.

Cost, Effort and Time to Delivery

Lastly, a key factor that must be considered is the cost, effort and time to delivery. The in-house solution would require a large amount of effort, because firstly, individual liaising must be done with each of the database owners to gain indexing rights, and later, the bulk of the application would then

have to be developed by SMU. As a consequence, this approach would have a long time to delivery. The actual monetary costs to build and maintain the system would depend on whether SMU outsources the work to another vendor, or decides to utilize existing staff or students. Maintenance will be difficult to manage, since database content provider may change data formats anytime which cannot be indexed the same way formerly.

Unlike the in-house solution, joining the Library Links program at Google Scholar is free and requires minimal effort. According to the Google FAQ page, many university libraries only need to set the configuration options in their existing link resolver and then email Google to join the program. Thus, this solution is the most optimal in terms of monetary costs, time and effort.

Lastly, the cost of a commercial meta-search package depends on which one is selected, and the desired results return (a ranked list would lead to higher cost due to more complex algorithm).. A moderate amount of time and effort would be required to compare and choose among the packages available, and then customize the software later on. As such, while this solution is not as fast and cheap as using Google Scholar, it is not as resource intensive and costly as the in-house approach either, since primarily, it is simply an aggregation application which does not executes crawling.

Conclusion

Analysis of the three approaches reveals that the in-house approach allows for fine-grained control. However, it is difficult to implement because the effort and amount of time required to develop and maintain the solution is extremely high. In contrast, joining the Links Program with Google Scholar is fast and free of charge, although the SMU library will not be able to make customizations; while purchasing a meta-search product and then tailoring it will require a moderate amount of resources. Copyright issues too, are basic and critical factors to consider when indexing the database's content in-house, since all benefits of an in-house search application would disintegrate if these issues manifest. These are, in our view, unjustifiable and unnecessary risks which can be transferred to Google Scholar or other meta-search vendors.

As such, we suggest that the library at SMU implement two options – join the Links Program with Google Scholar and purchase a meta-search product, as they are complementary and will enhance the user experience of searching for academic materials. It costs virtually nothing to join the Links Program with Google, and this approach could be used by students and faculty who would not want to restrict their search to the documents within SMU subscribed databases. Conversely, the meta-search would allow users to easily query multiple databases that have been subscribed by SMU, and it could potentially provide a similar user experience as that of the in-house approach. This approach is also more likely to be feasible, as it only involves customizing a commercial application rather than building the entire application from scratch, and therefore the time to delivery is shorter.

Students and faculty alike would thus be able to fulfill their need to retrieve academic materials more wholesomely with both approaches at minimum cost. As a world class academic repository with outstanding databases can only be as good as what users are able to find, our recommendations are aimed at streamlining the current process of document retrieval, by providing the users two methods of searching across databases, both of which will save them effort and time.

References

- ⁱ "Metasearch engine - Wikipedia, the free encyclopedia." http://en.wikipedia.org/wiki/Metasearch_engine (accessed October 14, 2007).
- ⁱⁱ "Ex Libris - Metalib - Overview." <http://www.exlibrisgroup.com/metalib.htm> (accessed October 14, 2007).
- ⁱⁱⁱ "Innovative Interfaces - Trusted Library Technology and Service." <http://www.iii.com/> (accessed October 14, 2007).
- ^{iv} "Auto-Graphics, Inc - Library Automation Solutions, Products & Services." <http://www4.auto-graphics.com/products/agentsearch/agentsearch.htm> (accessed October 14, 2007).
- ^v Tamar Sadeh. "Google Scholar Versus Metasearch Systems." *HEP Libraries Webzine*, 2006. <http://library.cern.ch/HEPLW/12/papers/1/> (accessed October 14, 2007).
- ^{vi} "Google Press Center: Product Descriptions." <http://www.google.com/press/descriptions.html> (accessed October 14, 2007).
- ^{vii} "Metasearch." <http://www.cs.dartmouth.edu/research/node8.html> (accessed October 15, 2007).
- ^{viii} "Google Scholar Support for Libraries." <http://scholar.google.com.sg/intl/en/scholar/libraries.html#start1> (accessed October 14, 2007).
- ^{ix} "Google Scholar." January 2006. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1324783> (accessed October 9, 2007)