Federated Search Engines: A Building Block for Participatory Library Service

Purnima Upadhyay

Abstract

This article provides a brief background on the current state of information search in this highly connected and pervasive internet age. It is followed by an introduction to the federated search engine and its definition. While this article assumes that the role played by federated search engine in searching disparate information has been now quite justified and established, it lays the emphasis on the capabilities of the federated search engines. It does so by listing the functionalities expected from a federated search engine, and higher level explanation of inner workings of the federated search engine. It also lists the advantages and shortcomings of federated search engines. This is followed by the selection criteria of a federated search engine and lastly a list of some of the popular federated search engines available today.

Keywords: Federated Search, Participatory Library Service

1. Introduction

In this day and age of digital revolution, we have come a long way since the early days of internet and World Wide Web. Internet connectivity has become so pervasive that terms such as "portal", "information gateway" are no longer esoteric. Users are consuming and producing large amounts of various types of digital data on daily basis that persist on the web.

In spite of greater connectivity and consumption of information sources, it is becoming increasingly difficult for the users to access the right information. While there are excellent choices available for searching the web, a large part of the web still remains unavailable simply because it is not accessible to the conventional search crawler software. In addition, there is now a proliferation of subject specific information repositories that are released as proprietary databases and leased, licensed or sold to corporations, academic and research organizations, enterprises, etc. It is a common practice in the research fraternity to search such databases for citations and references.

Such information databases are accessible only through their proprietary access mechanisms and are otherwise heterogeneous in terms of their content, format and schema. They also follow their own search methodologies, and present search results in their own format. Often, organizations license multiple such databases and have their personnel trained to work with each one of them. The task of sequential searching for information in such databases one by one and collating results is quite onerous because the availability of such information repositories has increased exponentially.

Federated Search, also known as distributed information retrieval, provides one uniform interface for searching information from multiple sources. Federated Search Software, also called Federated Search Engine, runs the federated search query against all the available data repositories, retrieves the search results, and compiles into one list, thus greatly facilitating information discovery by means of simplifying the search process.

Within the framework of building participatory library services, the ability to conduct search among distributed federation forms the backbone of the infrastructure that connects disparate sources of information and makes it available to the user. This distributed search technology provides users with simultaneous access to diverse sources of information through a unified interface. Thus it becomes an integral component of an Information Portal that enables the users to search multiple sources of information through one search query.

2. What is Federated Search?

The federated search software sits between the user and information sources as a discovery tool that allows user to find information quickly from multiple sources. At its core, federated search is a single interface that is capable of simultaneously searching multiple data repositories. Corporate or enterprise intra-nets, fee-based or licensed databases, library catalogues, internet resources, user specific digital storage repositories, etc. are typical data sources handled by federated search.

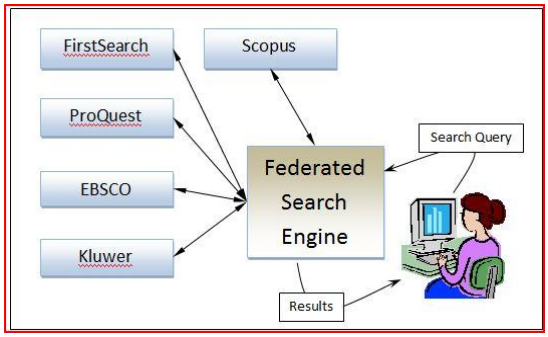

Figure -1

3. Definition

Within the domain on Library Sciences, federated search is defined as "Transforming a query and broadcasting it to a group of disparate databases with appropriate syntax, merging the results collected from the databases, presenting them in succinct and unified format with minimal duplication, and allowing the library patron to sort the merged results by various criteria"[1]. In simpler terms it can be defined as a search system that uses common interface to enable the user to simultaneously search data repositories from different providers of information. The user submits a single search query to the federated search engine, which presents the user with an aggregated search results obtained from all participating information provider.

4. Working of Federated Search Engine

The federated search engines work fundamentally different from the web crawler based conventional search engines such as Google, Yahoo! etc., also called surface web search engines. These surface web search engines follow a known set of links to discover new web pages. The newly discovered web pages also contain links, and in this manner they build up a link of links. While they are following

the links, they are also grabbing the contents of the web page and indexing it for search. This process is called web crawling.

The federated search engines (or also generically called deep web search engines) don't follow web crawling paradigm. Instead, the federated search engines search for the contents by programmatically interacting with the content sources using their published access methodologies. Each content owner provides its own mechanism for content search and retrieval from its data repositories. The federated search engine simply follows this access mechanism to search and retrieve the contents from providers.

A typical federated search engine has built in support to interact with various vendors' data repositories. Such capability of a federated search engine is implemented as what is called "connector" – a connector is what enables a federated search engine to interact with external databases and query and search for information.

Internally, the federated search engine is responsible for transforming the search query into individual search queries understandable by each of the information repositories, which are then executed in parallel by them. It uses appropriate connector to communicate the search queries to the data providers. The results received from the repositories are then merged into a single list before presenting it to the user. Following explains this process in more details.

4.1 Search Query Translation

Different search engines support different search syntaxes. Some require boolean operators (AND, OR, NOT) to be part of the search query, while others may not support it, and still others may consider them optional. Some may support wildcard searches, while others may not support them. Some engines may require string parameters to be enclosed in double quotes; some may require them enclosed in single quotes. The federated search engine accomplishes this task by employing appropriate connector to translate search queries.

4.2 Search Fields Mapping

If the federated search engine supports advanced search options, such as searching on multiple fields (author, title, etc.), then these fields have to be mapped to the target repository field naming system, if these fields are supported by the target repository. It may also be necessary to consider the semantics involved in the mapping process. For example, what federated search engine calls "Title" may be called "ArticleTitle" by the target repository.

4.3 Search Query Submission

This step could be as simple as submitting an HTML form – or it could be a complex process – especially if the search involves multiple steps. For example, some repositories may require to first performing a pre-search to see if there is any relevant information available. In such a case, the user may have to respond appropriately by providing further information to refine the search.

4.4 Search Results Retrieval

The format of the search results is determined by the information provider – it could be in the form of HTML (Hyper Text Markup Language) or XML (Extensible Markup Language), or any other proprietary format as specified by the information provider

4.5. Search Results Processing

Depending upon from where the information is retrieved, this step can be simple or complex. The federated search engine merges and collates the search results and processes them for the presentation to the user in a form that is easily understandable regardless of from where it was retrieved.

5. Features of Federated Search Engine

Typically, a federated search engine offers following set of functionalities. It is important to note that not all federated search software will offer all the functionalities. What follows is a comprehensive list of the features.

- ❖ Ability to search multiple databases concurrently
- ❖ Search databases in real time
- ❖ Simple and advanced search capabilities
- ❖ De-duping (de-duplication) of records – removing duplicates from the result set based upon a set of rules
- ❖ Merged search results
- ❖ Sorting of records
- ❖ Faceted (or clustered) search results
- ❖ Save, print or email search results
- ❖ Exporting of the search results to a file
- ❖ Extensive patron authentication support
- ❖ Display databases by categories
- ❖ Compliance with an OpenURL resolver
- ❖ Search status report
- ❖ Ability to search local and remote databases (using protocol Z39.50)
- ❖ Personalized access to resources
- ❖ Ability to access electronically available content without further authentication (one-time login)
- ❖ Relevance ranking of search results
- ❖ Unlimited simultaneous users
- ❖ Ability to link with inter library loan (ILL) system
- ❖ Extensive search statistics
- ❖ Customizable user interface (UI)
- ❖ Ability to execute distributed federated search in various sections of the web site (to reduce search load)

6. Advantages and Shortcomings

The federated search engines offer some obvious advantages:

1. A single search gateway into multiple and diverse sources of information – a "One-Stop-Searching".

2. Information discovery is enhanced and improved. With advances in the computer science field, the information search retrieval technology is constantly improving in terms of its performance and accuracy.

3. A very easy learning curve for the users. Users can search multiple sources without having to learn the search syntax for all of them.

4. Ability to search library as well as non-library content such as corporate intra-net, world wide web, etc.

5. Ability to further process the search results – such as save, print, sort, refine the search results.

These advantages come with certain shortcomings. Most of these shortcomings result from its

requirement to provide common and unified interface to search multiple datasources.

1. Federated search interface is not as detailed as native search. A native search interface for a particular datasource will always offer more options to the user.

2. Similarly, compared to a native search, a federated search may not be able to fully utilize search capabilities of target data repository due to its limited search interface.

3. Complete and correct de-duping may not be possible.

4. Any changes in the data repository schema, format or search protocol will require extensive changes to the federated search engine.

5. Being a federated datasource, there are possibilities that changes in the datasources will render it unavailable for the federated search due to change in the format, schema, protocol.

6. Federated search engine is best for discovering the information that is unknown. If the user knows what article he or she is looking for, a direct native search will be more effective and less time consuming.

7. Continuous maintenance of a federated search engine so that it functions efficiently and is able to interface with available datasources, is an affair that requires technical and financial resources.

7. Federated Search Engine Software Selection Criteria

There are many software vendors who are in the business of developing and selling federated search engines. Some of these are quite successful and popular, and some are new entrants. There are also many free open source software projects which too provide federated search engines. Some of these projects are stagnant and some are quite active.

Because investing in federating search engine assumes a certain level of commitments in terms of fiscal and technical resources, it is very important to be able to choose a right software and vendor.

The following list details some of the criteria in various categories to be considered in the selection process.

7.1 Usability

a. User friendliness of the user interface
b. Training required to learn to use the product
c. Web based help pages
d. Browser support
e. Guided search (wizards to define search query)
f. Easy navigation on the search and result pages
g. Clustering or other type of visualization of results
h. Search performance (fast, slow)

7.2 Software architecture

a. Number of sources that can be searched at a time
b. Number of simultaneous users of the system
c. Use of open source versus proprietary software components
d. Customizability of the internal algorithms (sorting, relevance ranking, etc.)
e. User authentications per source
f. User management (rights, access privileges, etc.)

7.3 Integration

a. Application Programming Interface (API) availability
b. URL resolver integration
c. Standards based web interface
d. Inter-library Loan Service (ILS) integration
e. Social networking web site integration

7.4. Search Features

a. Limit search to specific datasources
b. Search full text versus abstract
c. Parallel search in real time
d. Support for boolean operators and wildcards
e. Save Search Queries for later use
f. View history of searches
g. Limit search by date, geographic region, etc.

7.5 Search Results

a. Relevance ranking
b. Duplication removal (de-dupe)
c. Single integrated results page (vs. individual results page per each datasource queried)
d. Limit of the search results from a datasource
e. Incremental and partial results
f. Save or export search results

8. Federated Search Engine Products

There is numerous commercial and open source federated search engine software available. The following list includes some of the major players.

8.1 Commercial

a. WebFeat

This is one of the pioneering and early federated search software that is available commercially. It is very widely used in the academics and the government.

b. 360 Search

This federated search engine offers SaaS (Software as a service) model.

c. Muse

Other than federated search, this company also provides content integration, aggregation and transformation solutions.

d. Explorit

This federated search engine also permits integration with the intra-web using their programming interface.

2. Open Source

a. LibraryFind

This is developed by Oregon State University, Oregaon, USA

b. MasterKey

c. Sesat

Provides search middleware software with federated search capability

d. OpenSiteSearch

Implements Z39.50 portal system for library search application.

9. Conclusion

The federated search engines are very useful in searching information that is not crawled by the conventional search engines. There are many vendors who are in the business of selling and licensing proprietary information which they have gathered and compiled over many years. Federated search enables the consumers of such information sources to access them conveniently and reliably.

However, there are many vendors of the federated search technology, and not all of them provide similar level of functionality. With more and more federated search engine vendors adding support for accessing these information sources, it is again becoming possible that federated search engine technology may help create "one-stop search" for a particular domain of knowledge – if not all domains.

References

1. **Explorit**, available at web site http://www.deepwebtech.com/company/resource-center/explorit-features/ (accessed on 26/12/2011)
2. **Internet Insights – Thoughts about Federated Searching**, available at website http://www2.hawaii.edu/~jacso/extra/federated/federated.htm (accessed on 26/12/2011)
3. **LibraryFind**, available at web site http://www.libraryfind.org/ (accessed on 26/12/2011)
4. **MasterKey**, available at web site http://www.indexdata.com/masterkey (accessed on 26/12/2011)
5. **Muse**, available at web site http://www.museglobal.com/ (accessed on 26/12/2011)
6. **OpenSiteSearch**, available at web site http://opensitesearch.sourceforge.net/ (accessed on 26/12/2011)
7. **Serials Solution**, available at web site http://www.webfeat.org/ (accessed on 26/12/2011)
8. **Serials Solution**, available at web site http://www.serialssolutions.com/discovery/360-search/ (accessed on 26/12/2011)
9. **Sesta**, available at web site http://sesat.no/ (accessed on 26/12/2011)

About Author

Ms. Purnima Upadhyay, Assistant Professor, Department of Library and Information Science, Gujarat Vidyapith, Ahmedabad.