DESIDOC Journal of Library & Information Technology, Vol. 32, No. 5, September 2012, pp. 427-430

Open Source Software for Federated Search

Shamkant Deshmukh*, Sonia Bhavsar** and Sandeep Bhavsar***

**S.S. Patil Arts, Commerce and Science College, Jalgaon-425 001*
***H&GH Mansukhani Institute of Management, Ulhasnagar-421 003*
****Dr V.N. Bedekar Institute of Management, Thane-401 105*

ABSTRACT

The article explains the concept of federated search and demarcates the difference between federated search and other search engines. Advantages of federated search have been described with technologies used for federated searching. Article also covers open source software available for federated search and given some federated search applications of public domain.

Keywords: Open source software, federated search

1. INTRODUCTION

Federated search is the necessity of today's users. INFLIBNET has played the vital role in providing e-resources to universities and colleges through UGC-INFONET and NLIST. In the last decade major consortia have emerged with annual investment stakes ranging from Rs. 25 lakhs to Rs. 50 crores. INDEST is the consortium of IITs, IISc, IIMs and the Engineering Colleges funded by MHRD and AICTE.

The AICTE has made it mandatory to subscribe e-resources to its affiliated institutes from 2012-13. It shows that day by day subscription of e-resources will be more as compare to print. Because of all these initiatives even a small library is having good number of e-resources and now the challenge is to improve the usage statistics of e-resources. Federated search plays an important role in disseminating the information from all these databases in one search.A federated search is the simultaneous searching of multiple online databases, with the facility to see a list of returns from each source with clickable links that will connect directly to the source database.Jasco[1] defines federated search as, "Transforming a query and broadcasting it to a group of disparate databases with the appropriate syntax, merging the results collected from the databases, presenting them in a succinct and unified format with minimal duplication, and allowing the library patron to sort the merged result set by various criteria". In simple words, federated searching will be defined as a search system using a common interface that enables the simultaneous searching of databases from a variety of vendors. Federated search technology enables users to search multiple information resources simultaneously through one search query. Users can then view search results in a single integrated list. In other words, users do no longer need to consult each information resource individually. Instead, they can search multiple library catalogues (OPACs), websites (e.g. Amazon.com, Google, etc.), subscription and citation databases all at once.

Federated search technology is an integral component of an information portal, which provides the interface to diverse information resources. Once the user enters his or her search query in the search box of the Information Portal, the system uses federated search technology to send the search string to each resource that is incorporated into the portal. The individual information resources then send the information portal a list of results from the search query. Users can view the number of documents retrieved in each resource and link directly to each search result. People find no difference between federated searching and metasearch but there is slight difference between federated and meta searching. Some federated search options are shown in Table 1.

2. DIFFERENTIATING FEDERATED SEARCH FROM OTHER WEB SEARCH ENGINES

Federated search engines differ from web search engines such as Google in a number of ways:

Received 14 May 2012, accepted 14 May 2012, online published 7 September 2012

Table 1. Federated search options

Website address	Content searched
Mednar.com	Medical information sources
Biznar.com	Business-related sources
WorldWideScience.org	Science content from all over the world, from government agencies, as well as other quality research and academic organisations
http://search.smartlib-bibliogen.ca/zengine?VDXaction=ZSearchSimple	Capital Smar tLibrary Consortium of Libraries
http://osulibrary.oregonstate.edu/metafind/about.html	Oregon State University's Library
http://scienceroll.polymeta.com/search/ui7/searchfr.jsp?un=scienceroll	A medical student's journey inside genetics and medicine through Web 2.0
Science.gov	Science documents from a number of US federal government agencies
http://lifesearch.indexdata.dk/#	University of Copenhagen's Library of Faculty of Life Sciences
Scitopia.org	Digital libraries of leading science and technology societies
http://www.techxtra.ac.uk	Thirty one different collections relevant to engineering, mathematics and computing, including content from over 50 publishers and providers

(a) Access to Content

Web search engines do not have access to high-quality information that exists in secure knowledge bases. These data stores need to be accessed by federated search technologies. This is also true for businesses seeking a portal to their internal applications.

(b) Speed of searching

Web search engines use a technique called 'crawling' to search for relevant surface information that is readily available in the public domain. This information can be retrieved more quickly than using a federated search as the data is superficial and may or may not be relevant. The performance of federated search engines is dependent on the underlying data stores and their ability to perform. There are performance-tuning strategies available to tune the federated search engine.

(c) Relevancy of content

Content retrieved from web search engines may not be relevant, as the web engine only crawls surface data. Depending on when a page was last crawled, the results may be a week, or a month out of date. Federated search engines use their own relevancy search algorithms that ensure that results are meaningful and relevant. Searches are done in real-time, so searches will always return current information.

(d) Merging of and ranking content

Federated search engines and web search engines rank results based on their own sorting algorithms. Additionally, federated search engines can be configured to merge and remove duplicates during the ranking process.

3. ADVANTAGES OF FEDERATED SEARCH

There are certain advantages of using federated searches. Some of them are:

- Reduced time it takes to do a basic search
- Unified access to diverse content sources
- Simultaneous searching across all sources
- Ability to simple search as well as advanced search
- Integrated results which are easy to view and use
- Direct links to the native source for further searching
- Ability to filter, sort, save, print, export, and e-mail

4. FEDERATED SEARCH TECHNOLOGIES

There are mainly four technologies used for federated searching:

4.1 Screen Scrapping or HTTP

HTTP (Hyper Text Transfer Protocol) is the single most important technology that drives the web and yet remains virtually transparent. Without this protocol, HTML and XML via the web would not be able to perform the myriad of tasks that are put to them daily. HTTP is an application-level protocol for distributed, collaborative, hypermedia information systems. HTTP has been in use by the world wide web global information initiative since 1990. The HTTP protocol is a request/response protocol. A client sends a request to the server in the form of a request method, URI, and protocol version, followed by a MIME-like message containing request modifiers, client information, and possible body content over a connection

with a server. The HTTP communication usually takes place over TCP/IP connections. The TCP guarantees that packets arriving to and from the web server are error free and in the right order. It doesn't, however, guarantee that packets arrive no matter what the network conditions are. When communications are congested or unavailable webpage delivery is slow and can time-out.

4.2 Z39.50 Protocol

Z39.50 is an American national standard for information retrieval. It is formally known as ANSI/NISO Z39.50-1995 - Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. This document specifies a set of rules and procedures for the behaviour of two systems communicating for the purposes of database searching and information retrieval. As a network application standard, Z39.50 is an open standard that enables communication between systems that run on different hardware and use different software The Z39.50 standard was developed to overcome the problems associated with multiple databases searching such as having to know the unique menus, command language, and search procedures of each system accessed. Z39.50 simplifies the search process by making it possible for a searcher to use the familiar user interface of the local system to search both the local library catalogue as well as any remote database system that support the standard. In libraries, the Z39.50 protocol is most often used for searching OPAC sources. The important facilities offered by Z39.50 are as follows:

4.2 Browse

Allows the client to scan the contents of wordlists or indexes on the server. This can be particularly useful in the case of controlled keyword lists or facets.

- **Access and resource control:** Allows authentication of users, and cost control and online charging for commercial services.
- **Sort:** Allows the client to request different orderings of query results, e.g. relevance ranking, sorting by date or version number, etc.
- **Explain:** Allows the client to interrogate the server about a number of details about its contents and its level of support for the application profile.
- **Item order:** Allows offline ordering of materials in cases where they cannot be delivered electronically, or where per-unit charging (e.g. online charging) is required. Such services are being supplied in an ad-hoc fashion by online web-based component repositories such as ASSET. The item order service provides a ready-made and semantically standardised version of this service.
- **Item update:** Permits an authorised client to update the contents of the remote database.

4.3 Search/Retrieve Web Service

Search/Retrieve Web (SRW) Service is a new http-based information retrieval protocol providing broadly the same facilities as Z39.50, but by means of very different technology. The SRW is designed to be a low barrier to entry solution to performing searches and other information retrieval operations across the internet. It uses existing, well tested and easily available technologies such as SOAP and XPath to perform what has been done in the past using proprietary solutions. The protocol has two ways that it can be carried, either via SOAP or as parameters in a URL. This second form is called Search Retrieve by URL (SRU). Other transports would also be possible, for example simple XML over HTTP, but these are not defined by the current standard. The primary function of SRW is to allow a user to search a remote database of records. This is done via the search retrieve operation, in which the client sends a search retrieve request and the server responds with a search retrieve response. The request has several parameters, most of which are optional. The response is primarily a list of XML records which matched the search, along with the full count of how many records were matched.

4.4 XML (EXtensible Markup Language)

XML stands for EXtensible Markup Language. The XML is a markup language much like HTML and was designed to carry data, not to display data. XML tags are not predefined. One must define his own tags. XML is designed to be self-descriptive and it is recommended by the World Wide Web Consortium. It is a fee-free open standard. XML is not a replacement for HTML. HTML is about displaying information, while XML is about carrying information. In simple words, XML is a software and hardware independent tool for carrying information. It is used both to encode documents and serialise data. It supports unicode, allowing almost any information in any written human language to be communicated.

XML is now as important for the Web as HTML was to the foundation of the Web. XML is everywhere. It is the most common tool for data transmissions between all sorts of applications, and becomes more and more popular in the area of storing and describing information. The XML simplifies data sharing as in the real world; computer systems and databases contain data in incompatible formats.

The XML data is stored in plain text format. This provides a software and hardware-independent way of storing data. This makes it much easier to create data that different applications can share.

5. OPEN SOURCE SOFTWARE FOR FEDERATED SEARCH

Pazpar[2] is a middleware web service, which allows libraries to develop their own interface in the programming language of their choice. This requires significant development time. Index Data offers, for a fee, MasterKey, a hosted, fully-customised and configured federated search tool. dbWiz is a MySQL and Perl-based federated search tool. It is part of a larger suite of tools called reSearcher, which Simon Fraser provides for managing electronic resources, and works with Simon Fraser's Godot OpenURL resolver. LibraryFind is a MySQL, Ruby-on-Rails-based federated search tool. It can search Z39.50 databases, open archives initiative (OAI)-capable databases, and OpenSearch-capable data resources. Unlike many federated search tools, LibraryFind has a built-in API, which allows developers to create their own interface or use LibraryFind search results in unique ways. The software is also capable of querying the API of an OpenURL resolver; determining whether or not full text is available and creating a link directly to that full-text object.

6. CONCLUSIONS

Federated searching reduces the time it takes to search and usually displays results in a common format. Most complete federated search solutions support multiple search protocols. Typically they offer integrated OpenURL resolution, spell checking, saved searches, alerts, de-duping, and single click access to the native interface. Federated search truly not serve as one-stop shopping for all library databases as people hoped, because some databases cannot be searched by the federated search for technical limitations.

REFERENCES

1. Google. WebMaster Tools. Google basics: Indexing. 2009.
2. http://en.wikipedia.org/wiki/Federated_search
3. http://searchengineland.com/understanding-federated-link-building-a-primer-with-examples-21056
4. http://www.deepwebtech.com/
5. http://www.deepwebtech.com/PDFs/Federated%20Search%20Primer.pdf
6. http://www.libraryjournal.com/lj/ljinprintnetconnect/888213-335/karen_coombs_describes_the_university.html.csp
7. http://www.searchitright.com/federated-search-engine-article.asp
8. Lederman, Sol. Crawling vs deep web searching? Deep web technologies. Federated Search Blog, 17 December 2007. http://federatedsearchblog.com/2007/12/17/crawling-vs-deep-web-searching/
9. Collins, Maria D.D. & Carr, Patrick L. Managing the transition from print to electronic journals and resources: A guide for library and information professionals.