SEARCHING THE HEP WEB

Jeff J. Dingbaum

HEP Network Resource Center
Fermi National Accelerator Laboratory
Po Box 500, Batavia, IL 60510, USA

Finding the right information is often critical to the success of a project, whether it be researching a paper or finding the top quark. Since HEP researchers are often scattered widely around the globe it is hard to find the correct information, even from members of the same collaboration. With the advent of the world wide web as a large document repository, more and more small groups are putting their documents onto workgroup sized web servers. Realizing that locating the information contained in a document or even the document itself may be a time consuming endeavor, the HEP Network Resource Center at Fermilab has been indexing HEP world wide web servers. HEPNRC is indexing many types of HEP-related, web-based documents from many different web servers. By gathering indices nightly, searches often take no longer to complete than to display the results back to the user. Prototype efforts with collaborations such as Babar, US-CMS, DART, and SDSS have been very successful and operations are being expanded to include experiments, major universities and even large HEP laboritories still lacking in search technology. As a result of this project, someone can search several HEP sites all from one central site. Documents types now indexed by the HEPNRC spiders include text, html, and Adobe PDF.

1 Search the Global HEP Web

A problem that geographically separated experiments and collaborations commonly have is that of finding documents that have been put on the web. Larger sites may have a central document repository, while smaller sites may have a local web server with just their documentation. Finding documents from each of these types of sites has its own challenges. A larger site has many documents, but they can be difficult to locate if the person searching is not familiar with the file structure of that server. Smaller sites may offer excellent documents,

Preprint submitted to Elsevier Preprint *23 April 1997*

but people may pass by their web site because they do not know the URL to get there. HEPNRC is working to solve both problems by using full text indexing of HEP related sites. By indexing HEP WWW servers and maintaining them at a central HEP site (HEPIC - http://www.hep.net/), users will be able to search for documents on many different servers without having to find and visit each site. Because HEPNRC is indexing many HEP related sites, large and small, users looking for HEP data can start their search at the "GLOBAL HEP SEARCH" on HEPIC. Trial efforts with the BaBar collaboration have met with good results and has expanded into a production system that also includes Carnegie Mellon University, CMS (Compact Muon Solenoid), DAH (Data Acquisition Hardware), DART (Data Acquisition Project), Fermilab, SDSS (Sloan Digital Sky Survey), and Texas Tech University.
Search options prior to those at HEPIC were limited to commercial services such as Yahoo and AltaVista. While these types of services can be valuable, they tend to become out of date rather quickly. Because of the scope that they offer their frequency of indexing is less. A search robot from one of these types of sites may index a site and not come back for months, while the documents on the server may change substantially. This tends to make searches at these sites inconclusive at best. Conversely, using many of the home-grown searching tools, such as those using perlbased scripts do searching in real time, thus providing very poor response times, as well as returning only a limited amount of "hits". Also, many of these types of programs only return exact matches and are of limited value to the user. Using the Verity based HEPIC search pages, a user queries on indices that are updated nightly. HEPIC users can do full text searches on html and text files. Pdf files are now searchable on all collections hosted on HEPIC.

2 Getting Meaningful Search Results - Fast

Because getting immediate results from a search is a primary concern indices to most sites are built and are then updated nightly. This is a computing and bandwidth intensive operation for HEPNRC. For every site to be indexed a spider has to go out and make the determination of whether or not to retrieve a document. To do this it does an HTTP HEAD, that is, it retrieves only some information about the document, such as how old it is, etc. If the document has not been indexed into the collection the spider will get the entire document. If by comparing the head information with that stored in the Verity database, it is determined that the document hasn't changed, it will go on to the next document. Building the indices/collections at night takes advantage of the relative idleness of the indexing machine and the machine being indexed, as well as the network.
Users executing a search on HEPIC see almost immediate results because they

are actually searching on a collection of words contained on a web site and not the web site directly. It is important to note that the collection does not contain the actual documents, but rather a reference to the location of that document on the web site. In this way searches are performed against the list of words in the collection. When a match is found the search engine returns a pointer to that document. If a user chooses to search for several words in the document the search engine will look for all or any of the these words. The results list will be ordered such that documents containing numerous examples of all the words are placed near the top of the list, while those that have lesser numbers of those words are placed near the bottom.
Search queries on phrases are treated similarly. The search engine looks for the query words individually and looks for proximity to each other within the document by looking at the location ob each query word within the document. Times for these types of searches range anywhere from 10-300 milliseconds, depending on complexity of the query and size of the collection being searched. Conversely, searching on systems that use pattern matching in real time can take 20-30 seconds or more.

3 Indexing Your Site

Adding additional sites is fairly easy to do. When a new site wishes to have their site included on the HEP search page a few simple steps are performed. First, the site to be indexed is visited to see if there is enough information to warrant indexing. A site that has under a dozen pages does not really need a search engine. If it is determined that it is a good site to index then if that site has a logo or image associated with it a copy of it will be brought back to HEPIC.
The next step is to make an html page for your site. This is done by taking a standard template and modifying a few parameters such as the site name that goes into the title and the top level header. The logo that was copied from the original site is then inserted into the html page at the top left corner, such as in the Babar example (see figure 1). This way most collaborations or experiments will be able to link directly to the search page that was created for specifically for their site.
The final step is to schedule and create a collection for that site. The collection will be run immediately or scheduled for indexing at a later time, depending on the amount and type of data to be indexed. If it is a reasonably small site with mostly html documents it can be indexed right away with little impact on each site's network traffic. If it is a larger site with many documents, especially non-html type, then an appropriate time will be determined and scheduled. These larger sites with complex document types are usually indexed at a time when the impact of indexing will have little or no effect. Since HEPNRC is

based at Fermilab in the U.S. Central Time Zone, those sites in the United States are usually run late in the evening.
At the time of this writing no European sites have been indexed, although several larger sites and collaborations are certainly good candidates. Since the author has an account on a machine at CERN, tests have been run using the Verity spider software on a machine there. Unfortunately the spider does not understand how to write the collections into the AFS filespace that was available. HEPNRC has requested Verity to fix this to allow the use of AFS in building collections, but it is not supported will probably not be fixed. Unless arrangements can be made to allow writing into native file space, this approach will not work for indexing European sites from CERN.
This may impact the frequency of the index updates, as larger more diverse sites such as CERN and DESY would be done on weekends, especially initial builds. Incremental updates will go faster, since less network and CPU resources are used, and thus they may be done nightly. First attempts at building indices of smaller European sites will be done before the larger sites. A logical time for these builds is probably from 2:00 to 7:00GMT. This timeframe would be ideal because it would be after hours at both ends of the transatlantic link, thus probably speeding file transfer because the links would be lightly loaded. In any case, the first run will be heavily monitored for possible loops indicating that the spider is indexing the same data again and again.
Another factor in determining when and how often a web site is indexed is the number of different document types that it contains. Sites containing mostly html and text MIME types are relatively easy to index and have the fastest per document indexing time. Document types such as PDF, Microsoft Word, Wordperfect, Microsoft Excel, Framemaker, can be indexed as well, but the amount of time taken for each of these types of documents is considerably longer. Not only because the files themselves are larger, but also due to the fact that filtering software is used to glean the data.

4 The Babar Search Page

An example of a site using the search engine on HEPIC is the Babar collaboration (shown in figure 1). Someone searching for information on the Babar web site would go to this page, either through the Babar web server or by going to the HEPIC search screens and selecting the Babar link. The example shown is a representative sample. If available, sites will have a small image in the upper left corner and will have their name across the top. This enables that site's webmaster to advertise that page. The navigation bar just under the site name points to various searching links.
The "Other HEP Search Screens" link points to a page that has links to numerous other search screens similar to the Babar page shown. The "Detailed

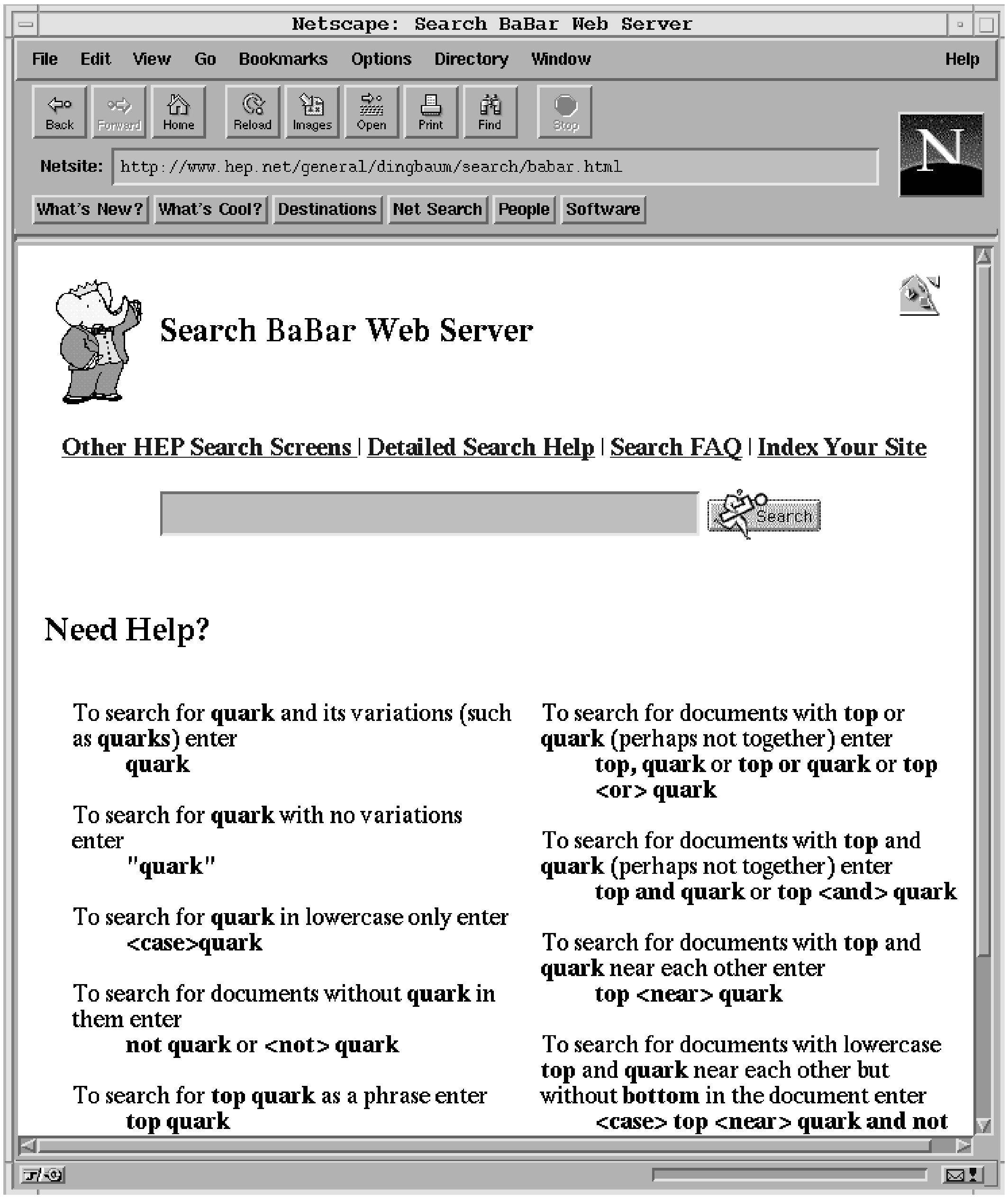

Fig. 1.

Search Help" is a link to a page on the Verity web server that goes into great detail about using the search interface. "Search FAQ" is a Frequently Asked Questions page that was compiled from users' questions about the search process. "Index Your Site" is a link to more information about getting the Verity software for use on your web server. The Verity software runs on a variety of platforms, but should your site not have a local machine that runs it, arrangements can be made for HEPNRC to index the web site remotely.

Below the navigation bar is the area where the user types the search criteria. Below that are some examples of how to use the search engine to your advantage. Users may make queries as simple or as complex as needed just by following the examples provided.

5 Partnering to Reduce Bandwidth

Nightly rebuilding of collections is fine for indexing a server or two on the local ethernet, but can tax a server and/or a network connection if done indiscriminately over the wide area. As stated earlier, great care is taken to monitor the collection building process, as well as to schedule collection building runs so that network traffic is kept to a minimum.
To alleviate network congestion and to assure that sites are indexed as often as they would like to be, HEPNRC has licensed the Verity search technology and can give copies of the indexers to other HEP sites at no cost. Each site can index their web server as often as they wish without congesting their off-site network connection. To use the index, it is first sent back to HEPIC using ftp, afs, or some other arranged method. Once it is placed over the old collection, the new documents added to the collection may be searched and they will show up in the results list.

6 Future Directions

HEPIC currently has over 10 sites that are indexed on a regular basis. Although most have been implemented without a problem, an occasional issue has arisen. Most notable is when the spider gets stuck in a loop and indexes a branch of a site over and over. This has happened on one occasion. It has not yet been determined if we can avoid this by directing the spider with more rigorous controls. Until this problem has been solved, careful monitoring of the spider will continue to be done.
The Global HEP Search project is looking for more sites, collaborations, and experiments to index. We are also looking for sites, particularly European, willing to cooperate with us and use the indexing tools to build collections locally at their site and then ftp the collections to us. Contact the author if your High Energy Physics related web site does not have search capability and you are interested in obtaining it. More sites are being added as requested and as time permits. For more information about the search project, or to add your web site to the HEPIC search pages, contact the author by email at dingbaum@hep.net or by phone at +1-630-840-8472.