Open Source System Implementation of Electronic Data Search over a Local Area Network

Ara Murad Vartanian (a

Abstract— despite the availability of commercial search software or library management systems, cost, setup or maintenance make them less viable for rapid deployment. The World Wide Web is not the only way to search for digital information. Local area network can be a viable alternative – and a necessary one in many operating environments. This paper describes an open source system for searching electronic content over a local area network. The system is easy enough to deploy quickly to handle the growing number of electronic resources. It is able to provide a convenient and simple search method based on Google's well known search engine technology. The system was implemented at the College of Pharmacy – University of Mustansiriyah and provides access to more than 42,000 electronic resources.

Index Terms— Search engine, LAN, DTS (Desktop Search), Google Desktop Search, Web Analysis, Open Source Software, e-Library.

1 Introduction

As electronic publishing matures, research and academic libraries are beginning to supplement their print holdings with electronic publications. This transition began with scientific journals, and is now advancing into academic and scholarly books. In the past few years, corporate and government libraries have also begun acquiring e-Books along with print holdings.

E-Books provide substantial advantages to libraries and their users. Both parties gain from 24/7 access, simultaneous user access, wider selection, and immediate updates, while libraries also benefit from back-end efficiencies, such as a lack of storage requirements, reduced maintenance costs, and reduced staffing time for physical handling and processing of printed books.[1]

Despite near-universal current practice, the World-Wide Web is by no means the only way to deliver digital library services; in many situations, universal access via the Internet is neither possible nor desirable. The Internet may suffer in terms of remote site availability and unpredictable network delays. In the developing countries, Internet connections are still either non-existent, undependable, or prohibitively expensive to use.[2]

2 Content Search

Since the PC era began in the 1980s, keeping data organized was relatively manageable due to the fixed small amount of storage available in the form of floppy disks. As the storage volume began to increase with hard disks becoming cheaper, the amount of files saved grew arithmetically until Internet use exploded in the 1990s. Up to that point, the user-friendly drag-and-click folder system to manage files was good enough. Folders organized logically made finding files easy. The system worked until the new millennium, when the Web transformed the computer into a giant storehouse of contacts, photographs, e-mails, music, videos, and all kinds of documents. The staggering volume of saved digital files made organizing file-and-folder system cumbersome, and searching for files started taking longer. What complicated matters is the way several programs – for e-mail, photo albums, contacts, and calendars – stored data. Getting the information meant having to run separate programs, each with its own search commands.

Google in 1998 made web search so fast that with a high-speed Internet connection, finding information in the Web was faster than looking for files on hard drives. Using the same technology that mapped the Internet's billions of pages, Google, Yahoo!, and MSN developed search engines programs to find files on a computer and are called Desktop Search Engines (DTS).

Desktop Search Engines are programs that help find the information people need by creating an in index of the contents of the files on hard disk drives (or parts of them) whenever the computer is idle. As each word in each file is indexed, the index file may grow to become very large. However, the payoff is that files are found almost immediately thus make the search for files fast and easy.

Desktop Search Engines not only make finding files faster, but makes managing files easier; as hard drives get bigger and prices go down, much more information may be saved without the need to delete files. Desktop Search Engines also reduce or eliminate the need to organize the information into folders since DTS engines can find files faster than navigating through folders. DTS engines also alleviate the need to memorize file names because keywords inside files are the elements used to conduct the searches.

As computers have become more connected, public data (especially in corporates, institutions or organizations) has become stored on mapped network drives and virtual folders. Some DTS engines are provided with the capability to search through mapped network drives or shared folders, but these programs are not free. Examples of the free desktop search engines are: Yahoo Desktop Search, Google Desktop Search, Copernic Desktop Search, Ask Jeeves, and MSN. Lakshmi et. al. [3] presented an objective comparison of the approaches

used by existing desktop search packages for crawling, indexing and search.

In a corporate network or an educational institution intranet, it is desired to provide search capabilities within the network rather than on separate machines. An obvious advantage is shown in the case of a digital library where electronic books, media or images are stored on a network drive and indexed by a single machine; although the index file will be created and updated on a single host, it will be used by many users across the network.

Huang [4] described a mechanism with integrated desktop search engines using a cross platform java application through the use of the UPnP (Universal Plug and Play) protocol. However, every computer in the intranet will install one of the desktop search engines and requires being equipped with both the UPnP server and the controller.

Google Search Appliance [5] is a rack-mounted device providing document indexing functionality that can be integrated into an intranet, document management system or web site using a Google search-like interface for end-user retrieval. It is supplied in three models: an entry-level appliance capable of indexing up to 300,000 documents (Google Mini), a 2U appliance capable of indexing up to 10 million documents, and a 5U appliance capable of indexing up to 30 million documents [6]. Unfortunately, these options are not free and the entry level Google Mini costs $3,000. This can be an expensive choice for educational institutes especially in the developed countries.

This paper describes the use of the freely available Google Desktop Search along with open source tools to return search results for users over a LAN through a web server. A Visual Basic parser is implemented to analyze the usage statistics by inspecting the web server log file.

3 PROPOSED SYSTEM

3.1 System Requirements

The system is comprised of two functional parts: a computer (server) that works as the repository for digital data and does the indexing and search, and client computers.

The server computer has the following components installed on it:

- Google Desktop Search (GDS) [7]: a free desktop application created by Google to provide indexing and searching capabilities of documents over a local computer. It works by indexing common file types (Microsoft Office documents, PDF files, multimedia files (images, sounds and video) and text-based files).
- Goolag [8]: a PHP application script that provides the capability to run Google Desktop Search (GDS) from a remote computer.
- Apache and PHP module: Apache [9] is an open source web server widely used for modern operating systems including UNIX and Windows due to its extensible and versatile features and its ease of configuration and deployment.
- TweakGDS [10]: is a freeware application used to modify some of GDS parameters. It is beneficiary to keep the index file on a separate disk drive as it is more efficient to access index data from a different drive rather than the same operating system disk. Also, in terms of disk space usage, moving the index files to a separate location eliminates the need to re-index the digital content in case of deployment on a different computer. It also helps to backup. it was shown that for first time indexing for (10,000) documents, it takes (30 minutes) to finish.
- Using VB.NET, a parser was written to read the Apache server web log files and convert them to a format compatible to use with WebLogExpert [11], a freeware log analyzer for Apache server.

Clients are the computers connected over the local area network and having a standard web browser installed on them

3.2 Implementation

Google Desktop Search v.5.9.1005 was used during the time of implementation. It is installed on a computer with a wired LAN connection running on Windows 7 operating system on an Intel i5 2.53GHz CPU and 4GB RAM.

Electronic books and presentations in the form of HTML, Microsoft Office files (DOC, PPT, and XLS), Portable Document Format (PDF) and media files (Video and Audio) were stored in a folder on a separate hard disk drive partitioned in NTFS format. More than 21 Gigabytes of data in more than 42,000 files constituted the digital repository. After GDS was installed, TweakGDS was used to change the index location of GDS to a partition different from the one where the system files existed on.

Apache web server version 2.2.4 for Windows is used, it was configured by editing its configuration "httpd.conf" file to specify the server IP and listening port.

Goolag PHP script is installed along with PHP runtime library on the server. The script configuration is modified as specified in its supporting documentation. Clients search requests are captured via the PHP script residing on Apache webserver and the results obtained from GDS are forwarded back to requestors. The search interface is shown in Figure 1.

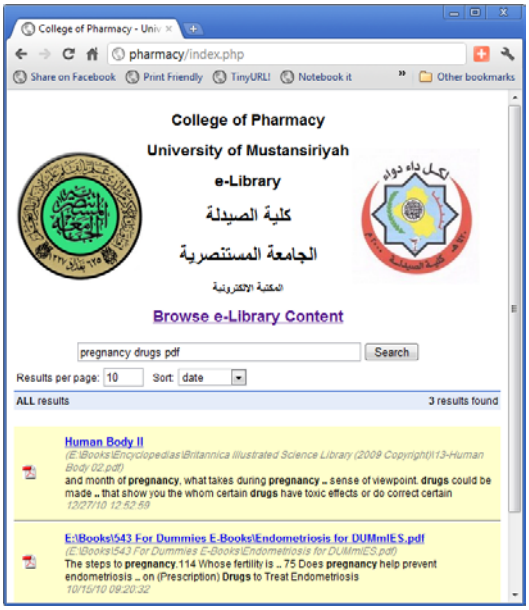

Fig. 1. Main Search Interface

A classical directory-browsing interface is included in the e-library main page as shown in Figure 2; this was implemented using Apache server. Although categorization of documents is not required, it is included to provide an easier human understanding for the placement of the files per specialty, field of interest or authorized personnel. This makes the addition of more content easier. It allows the user to sort the view based on file name, file modification date (which proves to be beneficial to locate recently added files), and file size.

Fig. 2. Sample Directory Browsing Interface

4 Experimental and Analysis

The initial index creation consumed about 4 hours of idle computer time; 606 MB of storage were needed for 20 GB data in more than 42,000 files. The size of the index file varies based on the content of files. 3% - 5% of the total data file is normal.

Since it is usually desirable to monitor the usage statistics of the system, the Apache web server was configured so that logging was enabled on it.

The web server log files were not directly suitable to display proper results when they were analyzed by log analyzers due to the fact that all the search requests were intercepted by the PHP Goolag script. A parser was coded in VB.NET to produce a log file format which contained the actual file requests that were embedded into the PHP script.

WebLog Expert Lite is used to analyze the parsed log files and activity statistics reports are generated for usage by days or by hours as shown in Figure 3.

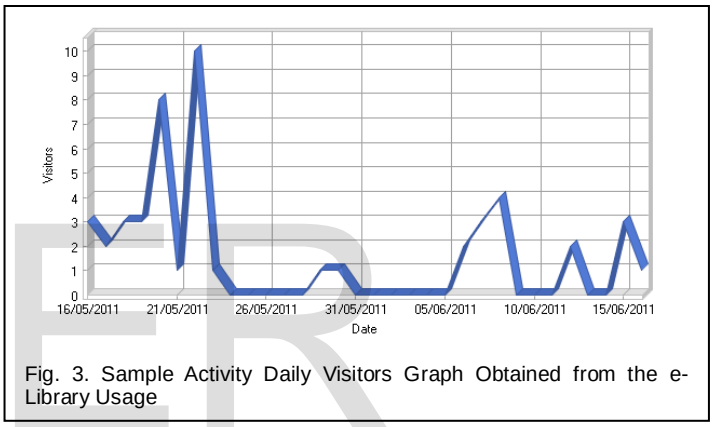

Fig. 3. Sample Activity Daily Visitors Graph Obtained from the e-Library Usage

The files which are downloaded the most can be viewed in order to monitor the active files that have the largest interest from users. Additionally, the number of users residing on the computers on the network is traced in addition to the total amount of data bandwidth consumed as shown in Table 1. The number of visitors is recorded. If a hit request from an IP address came after 30 minutes since the last request from this IP, it is considered to belong to a different visitor.

TABLE 1
SAMPLE COMPUTERS ACTIVITY FROM MULTIPLE HOSTS ON THE LAN

Host IP	*Hits*	*Visitors*	*Bandwidth (KB)*
127.0.0.1	690	47	128,057
192.168.0.79	210	6	5,475
192.168.0.210	46	5	154
192.168.1.251	24	2	48,226
Total	970	60	181,912

6 Conclusion

This paper presents a system for information retrieval without using the Internet. The components used in it are based on open-source software. The system does not require more

than a web browser at the clients' side while it maintains an updated index of the information that is being added. Usage statistics in terms of most searched items can point out to the interest trends and the number of users benefiting from the information over a period of time. Extra features like user authentication may be added to the system.

REFERENCES

[1] Rita A. Renner, Hoffman Marketing Communications, Inc. eBooks – Costs and Benefits to Academic and Research Libraries [White paper]. 2009. Retrieved from: http://www.springer.com/cda/content/document/cda_downloaddocument/eBook+White+Paper.pdf?SGWID=0-0-45-415198-0

[2] I. Witten, S. J. Cunningham, B. Rogers, R. McNab, and S. Boddie. Distributing digital libraries on the web, CD-ROMS, and intranets: Same information, same look-and-feel, different media. In J. Yen and C. C. Yang, editors, Proceedings of First Asia Digital Library Workshop: East Meets West, pages 98--105, 1998. Available: http://nzdl.sadl.uleth.ca/gsdl/collect/publicat/index/assoc/HASH0199/95cc7f7f.dir/doc.pdf

[3] Narasimhan, V. Lakshmi; Lowe, Michael. An Objective Comparison of Desktop Search and Visualization Tools. IEE, pp. 206-209. 2010. Retrieved from http://ieeexplore.ieee.org.tiger.sempertool.dk/stampPDF/getPDF.jsp?tp=&arnumber=05714640&isnumber=5714593&tag=1

[4] Huang, Wei-Lun , Lee, Tzao-Lin and Liao, Chiao-Szu, 'Desktop search in the intranet with integrated desktop search engines', in Computer Systems Architecture Conference, pp. 1-4. 2008.

[5] Google desktop web site, viewed 10 April, 2011, http://www.google.com/enterprise/search/

[6] Juan Carlos Perez, Google releases new version of its Search Appliance, viewed 2 June, 2011, http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9133839

[7] Google Desktop Search, retrieved on 20 Jan, 2011, http://desktop.google.com/features.html

[8] Goolag web site, viewed 10 April, 2011, http://www.asabox.com/goolag

[9] Apache Web site, viewed 12 April, 2011, http://www.apache.org

[10] Tweak GDS, retrieved on 20 Feb, 2011, http://www.podsync.com/software/TweakGDS_Setup.exe

[11] WeblogExpert Lite website, viewed 6 June, 2011, http://www.weblogexpert.com/