Zaval File Search
Version 1.3

User's Guide

Zaval Creative Engineering Group
http://www.zaval.org

Contents

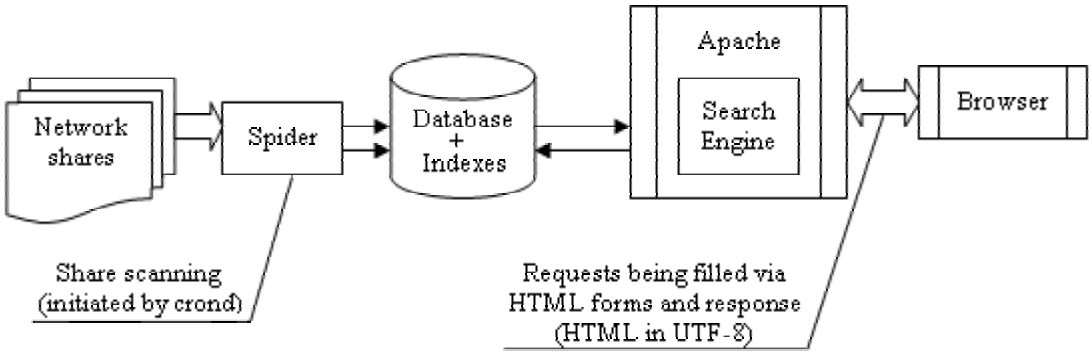
Introduction to Zaval File Search

The Zaval File Search tool is designed to provide easy and powerful indexing and search facilities in corporate networks with SMB/MS Network shares and FTP servers. Similar products are MS Indexing Service, and Napster-like tools.

Zaval File Search software has multi-tier client-server architecture, where

1. Back-end of the Zaval File Search is spider which collects all information asynchronously to all user requests;
2. Server-side is couple of scripts to produce fast search capabilities inside Apache web server;
3. Client side is any browser that recognizes HTML 3.2 or later. CSS and JavaScript support is optional.

In brief, architecture can be displayed as the following picture displays:



What Can You Do with Zaval File Search

The Zaval File Search provides all facilities to build an index based on SMB shares and FTP servers scan and than search through this index using user-friendly web based interface. It supports lots features like regular expression usage and search based on custom/predefined extensions.

Starting from v1.3.1 the Zaval File search engine allows incremental index building scheme, so if some hosts were turned off during scanning process they can be added to index later via incremental update (hosts that are already indexed will not be re-scanned to avoid network overload). This is really useful in large networks when you can't turn on every computer at the same time. The only thing you need to choose is scanning period. In almost all cases scanning period of one hour is enough to make relevant database with minor time of file links update.

When To Use Zaval File Search

The Zaval File Search best used for causal, irregular search for various files in your local network based on FTP and SMB file sharing solution. The powerful and flexible search engine and indexing service allow you to retrieve list of unique files have placed on the network via both MS Network and FTP shares.

Installing and Configuring the Zaval File Search

Requirements

In order to run Zaval File Search solution you need a Unix/Linux machine with the following tools installed:

- perl
- smbclient 2.0 or later
- standard ftp client
- Proper smb.conf settings, especially NLS options to support national symbols in share and file names
- Apache web-server

All these tools come within almost any Linux distribution, so probably you already have them installed.

Installation

The Zaval File Search currently distributed in three forms. There are Debian source packages, RPM-based source packages, and tarballs. All packages can be installed by standard and well-known package manager commands. The default settings in most cases can be leaved unchanged.

The Zaval File Search contains two parts:

- a batch of perl/shell/C-based search scripts and binaries that do actual network shares scan via smbclient and ftp tools and create indexes.
- perl-based CGI script that searches through indexes and displays results via web-based interface.

Note: both parts come as separate packages and can be downloaded from http://www.zaval.org/products/file-search/

The typical installation commands are

apt-get install zfsearch-spider

apt-get install zfsearch-client

or by packages

rpm -i zfsearch-spider-1.3.0.i386.rpm

rpm -i zfsearch-client-1.3.0.i386.rpm

or similar commands to the dpkg command. Sometimes you can get a message that several packages are missing - this can be when you have installed several packages from sources. If you sure that all packages are already installed you can use '--force' option for rpm.

Spider installation notes:

The package manager will register main script file in crontab to provide full-featured network scanning by default (usually default settings work fine). However, you can make necessary changes in configuration files.

Client installation notes:

The package manager will put 'search.pl' script and related files to the Apache's DocumentRoot in separate virtual host so all existing settings will be leaved unchanged. Make sure you have CGI.pm and mod_perl installed to spawn 'search.pl' script. This script does not write to any files.

Alternatively, you can build proper packages from corresponding src.rpm files as the rpm/make commands describes.

There are preferred ways to install Zaval File Search tools. But you can compile it manually from source tree by the following commands:

make

make install

make install_docs

This way is not recommended.

Configuring spider

If you decided to make changes make sure you set proper timeout value – an interval between two nearest network scans, because large network (more than 500 computers) scan requires several hours to complete. The exact value depends on the number of files on the shares and local network speed.

Note: The script tries to scan several hosts at one time, make sure there is enough free space and RAM on the host to avoid possible problems.

You can change all paths and options in **spider.conf** file. Make sure the appropriate client part have the same settings inside perl scripts too if you've changed the default settings. The example is listed below:

```
WINS=moon

DATADIR=/var/spool/zfsearch

HOST_LIST= /etc/zfsearch/hosts
EXCLUDE_SHARES=/etc/netscan/smb_ignore_shares

FTP_HOST_LIST=/etc/zfsearch/ftp_hosts
FTP_USER=ftp
FTP_PASSWORD=ftp@aol.com

LANG=ru_RU.UTF-8
```

```
TRANSLATE_COMMAND_SMB="iconv -f koi8-r -t UTF-8"
TRANSLATE_COMMAND_FTP="iconv -f CP1251 -t UTF-8"

TIME_RESCAN=12
TIME_ALLOW=72
```

where:

- WINS option is a network name of your network WINS server. This option allows retrieving list of workgroups, domains and appropriate master hosts. In theory you can specify any host in your network here, but choice of the dedicated WINS server (if available) is recommended. This option can be ignored if you have a list of hosts to scan.
- DATADIR option points to the temporary directory (used when spider builds indexes). Make sure you have enough free space available.
- HOST_LIST is a reference to file with network hosts listed in the following format:

  ```
  DOMAIN HOST
  DOMAIN HOST
      … …
  ```

  If the specified file is not available, WINS server will be used instead to build hosts list dynamically.
- EXCLUDE_SHARES specifies file with shares' names which won't be indexed in all cases. Regular expressions usage is allowed here.
- FTP_USER/FTP_PASSWORD allows specifying ftp user credentials to scan ftp shares.
- FTP_HOST_LIST specifies lost of ftp hosts. If this file is not specified or non-existent, the FTP scan does not performed instead of SMB scanning. Please use 3rd party network scanners to fill this file automatically.
- LANG is language settings for stored file lists. Make sure you have specified your locale correctly.
- TRANSLATE_COMMAND_SMB is a command for translate samba filenames list into UTF-8 encoding.
- TRANSLATE_COMMAND_FTP is a command for translate ftp file names list into UTF-8 encoding.
- TIME_RESCAN/TIME_ALLOW are time interval values (in hours).

There is also **smbclient.conf** file where you should specify domain options. See example below:

```
username=search
password=searchpwd
domain=workgroup
```

where

- username/password options are domain or workgroup logon credentials. You cannot leave these options empty. The specific user credentials to perform network scan are strongly recommended to manage access to server shares. For example, you can create a special user for this operation and than anybody who wants his/her shares to be available through search engine can add this user to 'read-only' list to the appropriate shares. This approach allows any shares that have private documents make them inaccessible for search engine's spider.
- domain is a local domain name.

Configuring client

There is only one thing you need to change: IP-address of virtual host in **/etc/zfsearch/zfsearch-httpd.conf** file (you need to change IP 127.0.0.1 to your real settings). After this operation include this file to your Apache configuration file (httpd.conf) with "Include" directive.

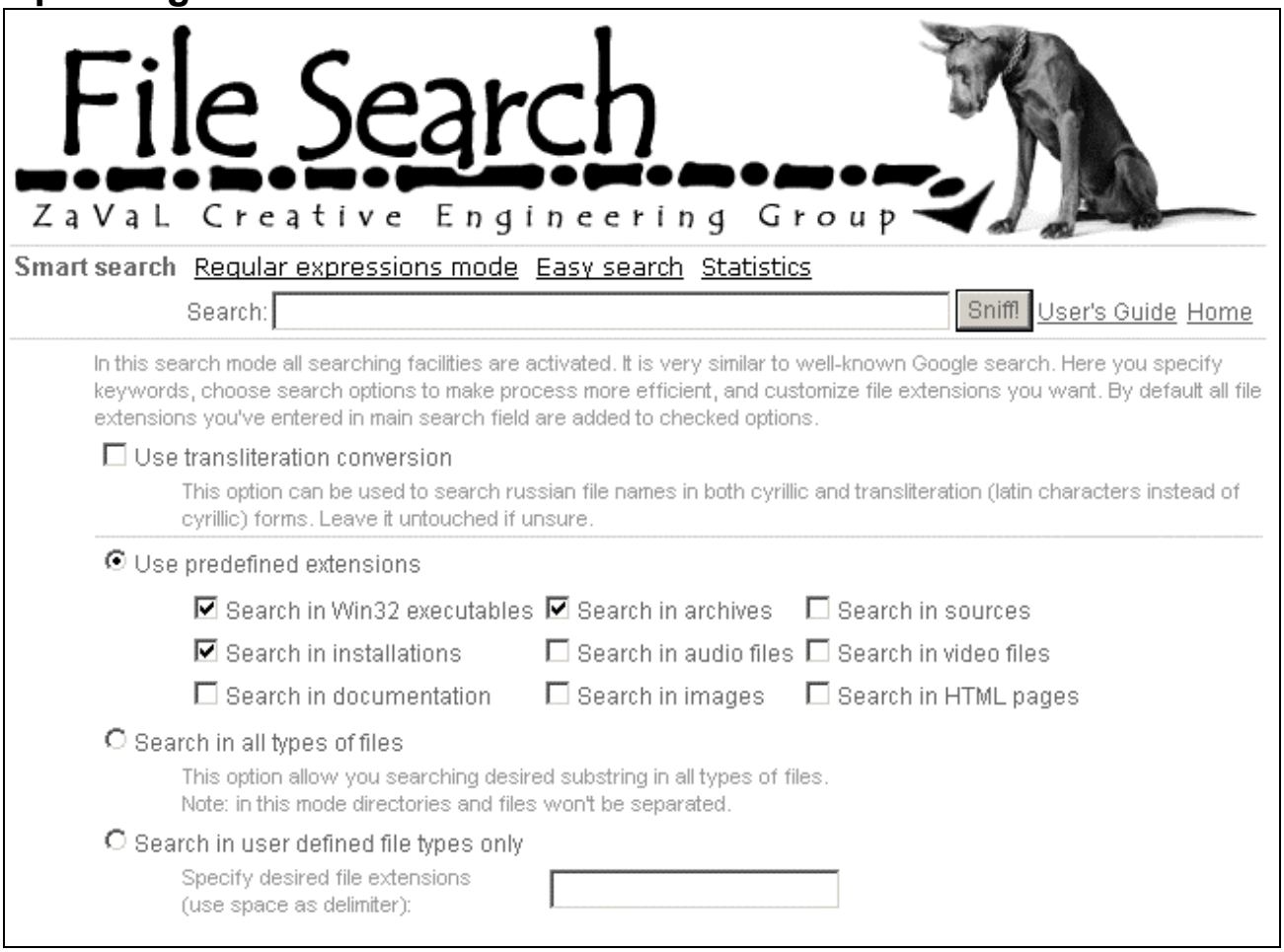
Example:

Include "/etc/zfsearch/zfsearch-httpd.conf"

All other settings can be leaved unchanged.

You can change all paths, options and SMB login in spider.conf file, however, make sure **/etc/zfsearch/client.conf** file have the same settings too.

Operating the Zaval File Search



To use the Zaval File Search you need a browser (it can be even lynx). All JavaScript code can be ignored in all cases, it was used to provide user-friendlier interface only (currently there is 'enable/disable' behavior only). Specify keywords to search for, choose appropriate settings and use 'Sniff!' button to do the search.

Note: you have to turn on "Always send URLs as UTF-8" in your browser to work correctly with requests in language other than English. In MS Explorer this feature can be found in *Internet options -> Advanced*.

The actual time you'll be waiting for the results depends on the number of files on the shares at you local network and the settings you have chosen. For a network about 200-300 computers this time can be about 2 seconds for reasonable requests on PII-450/128M. Better computer is able to operate much faster.

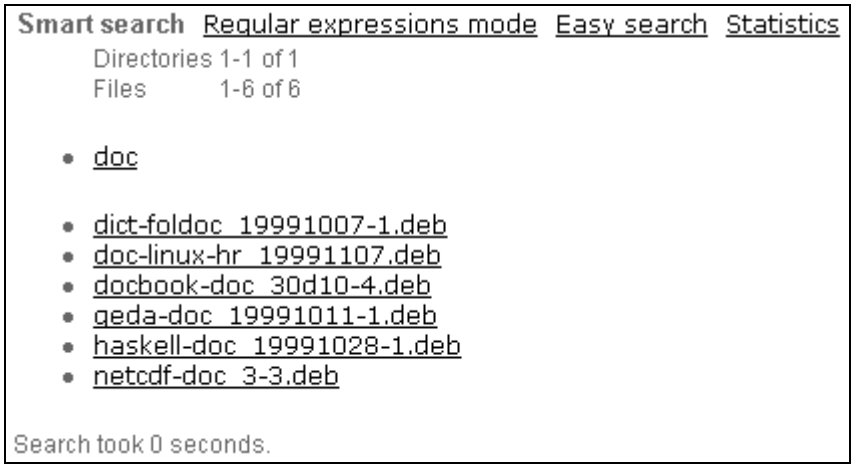
Zaval File Search has its own command syntax, similar to well-known Internet search engines, including grouping and logical operations. The regular expressions in perl style (without modifiers) and easy mode similar to well-known DOS meta characters are also applicable.

All wildcards, options and custom file extensions have accumulated to request if possible. The all-exclusive options provide enabling/disabling behavior in modern

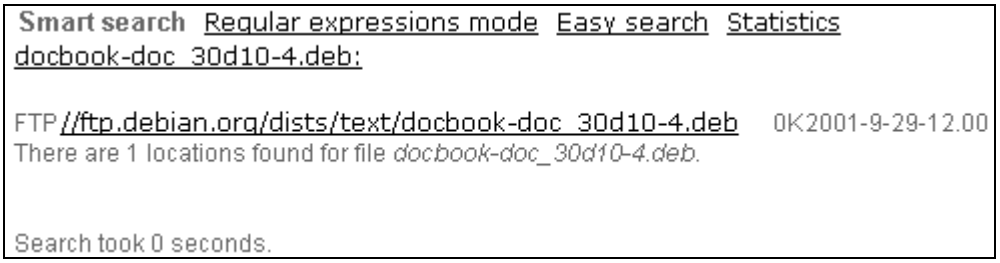
browsers such as IE/Mozilla to make option manipulations user-friendly. Use appropriate options and modes below to make request precise.

Zaval File Search: Search process notes

Zaval File Search provides non-standard search capabilities. We've decided to divide search process to the two important stages. There are search of file names without share names, full paths, sizes and other attributes, to make possibility to enumerate only unique names (here we tell "names", because we do not display unique files here); the second stage is display all files relevant to appropriate name being selected.



Screenshot 1. First search stage



Screenshot 2. Second search stage

The design of the search process briefly described above allows providing fastest search for advanced users. Some names and names wildcards are very popular, so the any files corresponding to one name will get a huge list of locations. Authors think the list of dozens of thousands files with the same names but with different places is not an effective way to manage and search files in network; few list with same names is better.

This feature allows users to use non-strict and informal requests for file names with wildcards to provide fastest navigation through the database. The logical operations bring additional flexibility in requests, so users can specify logical operations to wildcards in one request. In additional to two dimensional search Zaval File Search engine allow providing relevant but effective search process.

Zaval File Search Command Syntax

The following syntax constructions supported:

- expr ::= expr AND expr
- expr ::= expr OR expr
- expr ::= NOT expr
- expr ::= (expr)
- expr ::= <one word>
- expr ::= (<one word>|<quoted string>)+
- expr ::= <quoted string>

These constructions are rolled up to regular expressions by the following rules:

<quoted string>	All quotes are removed, all spaces are replaced to '\w', all '$', '/' and '@" symbols are replaced to escaped ones, the text is escaped by \Q \E brackets.
<one word>	All '$' and '@" symbols are replaced to escaped ones, the text is escaped by \Q \E brackets.
(<one word>|<quoted string>)+	All words and strings are replaced to the same regular expression construction as it declared in RBNF, and all elements of are conversed by two previous rules.
expr AND expr	This operation treated as simple element list (see previous).
(expr)	Simple grouping operation, equal to appropriate regex behavior.
expr OR expr	Logical 'or' operation, treated as (expr|expr).
NOT expr	Logical 'not' operation. Currently is not implemented.

Zaval File Search Modes and Options

The global modes are:

Smart search	In this search mode all searching facilities are activated. It is very similar to well-known Google search. Here you specify keywords, choose search options to make process more efficient, and customize file extensions you want. By default all file extensions you've entered in main search

	field are added to checked options.
Regular expressions mode	This option allows specifying Perl-like regular expression manually for advanced users and programmers. All strings be will used 'as is' with '$', '/' and '@" escaping as expression in the following construction: `print if ( /<your text here>/i )`
Easy search	This option allows specifying DOS-like wildcards to search. This mode was designed for inexperienced Windows users; they have no knowledge about powerful constructions. The easy search operates the only three forms of wildcards. There are '*' as replacement of any character sequence, '?' as replacement of any one symbol, and '[]' as one character of the list specified in square braces.
Statistics	This is not a search mode – you simply see list of hosts that are scanned right now with total number of files and total size for each of them.

In Smart search mode you can use several options to make your requests more accurate and precise:

- you can choose "Use predefined extensions" mode

Search in Win32 executables	The following file extensions will be used if this option is checked: exe, dll, com, ocx, bin, drv, sys.
Search in installations	The following file extensions will be used if this option is checked: exe, zip, tar.gz, cab, tar.bz2, bin, sh, deb, rpm.
Search in documentation	The following file extensions will be used if this option is checked: doc, rtf, txt, html, htm, htx, chi, chm.
Search in archives	The following file extensions will be used if this option is checked: arj, ain, arc, lha, rar, hqx, zip, cpz, tgz, gz, bz2, uc2, cab.
Search in audio files	The following file extensions will be used if this option is checked: au, aiff, wav, raw, mp3, stm, s3m, xm, mod, mid, midi, rmi.

Search in image files	The following file extensions will be used if this option is checked: jpg, jpeg, xpm, xbm, bmp, tif, gif, png, pcx, wmf, eps, tiff.
Search in sources	The following file extensions will be used if this option is checked: c, cc, cpp, cxx, h, hpp, java, jj, jack, pas, dpr, dfm, asm, asi, pl, sh, tcl, py, pm, l, y, lsp, 4, frt, bas, prg, rc, rh.
Search in video files	The following file extensions will be used if this option is checked: avi, mpg, mpeg, mpeg2, dat, dv.
Search in HTML files	The following file extensions will be used if this option is checked: htm, html, jsp, asp, asa, xtp, xsl, xml, dtd, xsd.

- Search in all types of files

  This option removes extension checks. Note: the search related to all files will produce a lot of files similar to. This option disables files and directory separation, and almost all heuristic algorithms have not used.

- Specify custom extensions

  The user-defined extensions can be specified here via separator (white space, comma, and semicolon) without leading 'dot'. If you specify '.rtf' extension the Zaval File Search will try to find file with '..rtf' extension.

All these options are limitations on the files' types to search through. All of them are translated to the following construction in regex:

```
<your command>(.)*\.(<ext1>|<ext2>|...)$
```

Regular expressions usage

This topic describes the syntax of regular expressions in search engine. All perlre (1) documented features are applicable.

Matching operations can have various modifiers. Modifiers that relate to the interpretation of the regular expression inside are listed below. The following options are used in all cases:

- do case-insensitive pattern matching.
- \Q and \E escaping (except 'allow reg. exprs' mode).
- $, \ and @ backslash escaping (except 'allow reg. exprs' mode).

The patterns used in search engine are the same as Perl pattern matching derive from supplied in the Version 8 regex routines. See appropriate documentation for details.

In particular the following met characters have their standard meanings:

\	Quote the next metacharacter
^	Match the beginning of the line
.	Match any character (except newline)
$	Match the end of the line (or before newline at the end)
|	Alternation
()	Grouping
[]	Character class

By default, the "^" character is guaranteed to match only the beginning of the string (file name), the "$" character only the end (or before the newline at the end).

The following standard quantifiers are recognized:

*	Match 0 or more times
+	Match 1 or more times
?	Match 1 or 0 times
{n}	Match exactly n times
{n,}	Match at least n times
{n,m}	Match at least n but not more than m times

The "*" modifier is equivalent to `{0,}', the "+" modifier to `{1,}', and the "?" modifier to `{0,1}'.

A quantified sub-pattern is "greedy", that is, it will match as many times as possible (given a particular starting location) while still allowing the rest of the pattern to match. If you want it to match the minimum number of times possible, follow the quantifier with a "?". Note that the meanings don't change, just the "greediness":

*?	Match 0 or more times
+?	Match 1 or more times
??	Match 1 or 0 times
{n}?	Match exactly n times
{n,}?	Match at least n times
{n,m}?	Match at least n but not more than m times

In addition, Perl defines the following:

\w	Match a "word" character (alphanumeric

	plus "_")
\W	Match a non-word character
\s	Match a whitespace character
\S	Match a non-whitespace character
\d	Match a digit character
\D	Match a non-digit character

A `\w' matches a single alphanumeric character, not a whole word. Use `\w+' to match a string of Perl-identifier characters (which isn't the same as matching an English word). If `use locale' is in effect, the list of alphabetic characters generated by `\w' is taken from the current locale. See the perl'a locale man page. You may use `\w', `\W', `\s', `\S', `\d', and `\D' within character classes, but if you try to use them as endpoints of a range, that's not a range, the "-" is understood literally.

The POSIX character class syntax is supported also. See perlre documentation for details.

Product limitations

This product was designed for irregular usage in small and medium Intranet space, and the engine was not optimized to obtain hundreds of million files in database. The engine was tested in Intranet with 300-500 computers with a lot of shares available (approximately 500 000 - 1 000 000 files in indexes only). The large database with thousands of available shares can cause significant slowdown when several users will be searching something like "all files containing 'a' symbol" at one time, especially if host computer is not very fast), so if you are seeking for a reliable Internet-related search engine (such as filez.com has) our engine is not for you.

However, it goes ok for a small company with 200-300 computers. ;)

Further product plans

Current tool implementation follows the minimalist-computing concept. The following features probably will be added:

- Logical 'not' operation;
- Additional file sources search such as NFS support in one box;

Support available

All support for software installation and problems should be sent directly to support@zaval.org with "Re: Zaval File Search Support" in subject line and plain text in the message body, describing your request and/or your problem. Since this software is distributed under the General Public License and is maintained by its authors on non-commercial basis, your request will be answered as soon as possible, but no later than 5 business days.

The Zaval Creative Engineering Group carries out its software customization/new software development on the regular basis. For more info contact us at info@zaval.org.