

NAME

index – indexing program

SYNOPSIS

index [-a] [-m] [-o] [-n *num*] [-q] [-N *num*] [-s *status*] [-t *tag*] [-u *pattern*] [-r *file*] [-g *file*] [*configfile*]

index -i [-u *url* | -f *file*] [-r *file*] [-g *file*] [*configfile*]

index -T *url* [-A *num*] [-r *file*] [-g *file*] [*configfile*]

index -S [-s *status*] [-t *tag*] [-u *pattern*] [-r *file*] [*configfile*]

index -M [-a] [-m] [-q] [-N *num*] [-r *file*] [-g *file*] [*configfile*]

index -C [-w] [-s *status*] [-t *tag*] [-u *pattern*] [*configfile*]

index -E | -D | -B | -K | -U | [-r *file*] [-g *file*] [*configfile*]

index -X1 | -X2 | -H [-g *file*] [*configfile*]

index -A *num* -u *pattern* [*configfile*]

index -P *URL* [*configfile*]

index -h | -?

DESCRIPTION

index is a component of **ASPseek** that performs Web crawling, documents downloading, parsing and storing. It can also be used to manipulate the **ASPseek** database.

During indexing process, **index** walks across the sites and stores found pages in a special data structures called delta files, and in SQL database.

When there is no more pages to index (or upon executing **index -D**), it sorts delta files and merges information from delta files into searchable database).

index supports HTTP and HTTP over SSL (https) protocols, and can parse documents in HTML and plain text formats. Support for other formats can be achieved via external converter programs.

The operation of **index** is mostly controlled by its configuration file **aspseek.conf(5)**, which is read upon startup. You can give configuration file name as a last argument to **index**.

OPTIONS

Indexing options

-n *number*

Index only *number* of documents and exit. Note that you should run **index -D** manually after running **index -n**. Actual number of documents indexed can be a little higher than value requested if you use many threads.

-N *number*

Run *number* of **index** threads. It makes sense if you have many different sites to index, since no two threads are allowed to index the same site.

-R *number*

Run *number* of resolver processes. Default is (argument of **-N** option)/5 + 1. It makes sense to increase the default value if your name server is slow.

-i Insert new URLs to database. URLs to insert can be given using **-u** or **-f** options.

Re-indexing control

-a Re-index all documents regardless of their expiration status. Normally (without this option) only documents that have indexed earlier than **Period** time ago are re-indexed.

-m Store words and hrefs found in documents regardless of their modification status. Normally (without this option) only those documents that have changed since last re-indexing are parsed.

-o Index documents with less hops first. Here "hops" means the "depth" value of the document.

-q Don't add URLs from **Server** configuration command (and their corresponding **robots.txt** URLs) to database. This can be used if you haven't changed your **aspseek.conf(5)** after last **index** run and is believed to speed up **index** startup in case you have several thousands **Server** entries in config.

- M** Index URLs which were indexed by previous indexing session. These URLs are stored in **tmpurl** SQL table. Used mostly for debugging purposes.

Indexing to real-time database

-T *URL*

Index *URL* to real-time database, so it will be available for searching in seconds. Note that you can't add too many documents to real-time database, otherwise the subsequent indexing to real-time database will be extremely slow. Actual limit of documents in real-time database is hardware dependent; well, about 1000 URLs should work OK. Documents from real-time database are merged to main database upon executing **index -D**.

This option is used to frequently re-index ever-changing pages (like first pages of news sites), or to re-index URL out-of-the-order (when you know it has just been changed) and see results immediately. Note that you can use **-A** option together with this one.

Clearing the database

- C** Clear the database. You can use subsection control options (described below) to limit clearing to some part of the database. Note that clearing with limits may be quite slow on large database.
- w** Used together with **-C** to disallow asking for confirmation before clearing.

Statistics

- S** Print simple database statistics. You can use subsection control options (described below) together with this option.

Subsection control

In most cases you can combine any of **-u**, **-s** and **-t** options.

-s *status*

Limit index to documents matching *status* (HTTP Status code, or 0 for documents that were not yet indexed).

- t** *tag* Limit index to documents matching *tag*. Tags can be set in **aspseek.conf(5)** file.

-u *pattern*

Limit index to documents with URLs matching *pattern* (supports SQL LIKE wildcard characters '%' and '_').

- f** *file* Read URLs to be indexed/inserted/cleared from *file*. You can use **-** as file name, in that case URL list will be read from **stdin**.

Output

- r** *file* Redirect output to *file*.
- g** *file* Sets indexing statistics log file name to *file*. Default is `/usr/local/aspseek/var/DBName/logs.txt`.

Stopping index

- E** Safely stop already running **index** process. Usable from scripts.

Database repairing

- X1** Check the inverted index for URLs for which **deleted** field in **urlword** SQL table is non-zero, or **status** field is not 200, or **origin** field is not 1.
- X2** Fix the above case by appending information about deleted keys to delta files. So, if you want to remove such records, run **index -X2**, **index -D** and finally perform SQL statements to delete unnecessary records.
- H** Recreate citation indexes and ranks file from **urlwordsNN.hrefs** fields in case of citation index corruption.

Database operations

- D** Merge delta files into main database. This implies **-B**, **-K** and **-U**.
- B** Generate subsets and spaces.
- K** Calculate PageRanks.

-U Calculate total number of non-empty URL, which is saved to /usr/local/aspseek/var/*DBName*/total file).

Miscellaneous

-P *URL* Prints path to specified *URL*. Here path means the way by which index found that URL by outgoing links.

-A *space_id*

Add/delete a site to/from web space (use together with **-u** or **-A** options).

Getting help

-h, -? Print short help page.

FILES

/usr/local/aspseek/etc/aspseek.conf

/usr/local/aspseek/var/*DBName*/logs.txt

SEE ALSO

aspseek(7), **aspseek.conf(5)**, **aspseek-sql(5)**.

AUTHORS

Copyright (C) 2000, 2001, 2002 by SWsoft.

Man page by Kir Kolyshkin <kir@asplinux.ru>