

NAME

aspseek-sql - the structure of SQL database tables used by ASPseek

SQL TABLES

wordurl

This table keeps information about each word in main and real-time database, one record per word.

Field	Description
<i>word</i>	Word itself.
<i>word_id</i>	Numeric ID of word .
<i>urls</i>	Information about sites and urls, in which word is encountered. Empty if size of info is greater than 1000 bytes, in this case info is stored in separate file.
<i>urlcount</i>	Number of URLs in which word is encountered.
<i>totalcount</i>	Total count of this word in all URLs.

Last 3 fields are used only if **CompactStorage** is set to *no*, and updated after finishing of crawling, or then **index(1)** is run with **-D** option.

wordurl1

This table keeps all information about each word in real-time database, one record per word.

Field	Description
<i>word</i>	Word itself.
<i>word_id</i>	Numeric ID of word , refers to wordurl.word .
<i>urls</i>	Information about sites and urls in which word is encountered. Always not empty regardless of size.
<i>urlcount</i>	Number of URLs in which word is encountered.
<i>totalcount</i>	Total count of this word in all URLs.

Last 3 fields are updated immediately after downloading of the URL by **index(1)** when it is run with **-T** option.

urlword

This table keeps information about all encountered URLs, both indexed and not indexed yet which match specified conditions in configuration files.

Field	Description
<i>url_id</i>	ID of URL.
<i>site_id</i>	ID of site, refers to sites.site_id .
<i>deleted</i>	Set to 1 if server returned 404 error and DeleteBad is set to <i>yes</i> , or if robots.txt or configuration rules disallow to index this URL.
<i>url</i>	URL itself.
<i>next_index_time</i>	Time of next indexing in seconds from UNIX epoch.
<i>status</i>	HTTP status returned by server or 0 if document has not been indexed yet.
<i>crc</i>	MD5 checksum of document.
<i>last_modified</i>	"Last-Modified" HTTP header returned by HTTP server.
<i>etag</i>	"ETag" header returned by HTTP server.
<i>last_index_time</i>	Time of last indexing in seconds from UNIX epoch.
<i>referrer</i>	ID of URL which first referred this URL.
<i>tag</i>	Arbitrary tag.
<i>hops</i>	Depth of URL in hyperlink tree.
<i>redir</i>	URL ID, where current URL is redirected or 0 if this URL is not redirected.
<i>origin</i>	URL ID of document which is origin of this cloned document, or zero if this is not clone.

urlwords*NN* (where *NN* is 2-digit number from 00-15)

These tables contain additional info about existing indexed URLs. Number *NN* in table name is $URL_ID \bmod 16$.

Field	Description
<i>deleted</i>	Set to 1 if server returned 404 error and DeleteBad is set to <i>yes</i> , or if robots.txt or configuration rules disallow to index this URL.
<i>wordcount</i>	Count of unique words in the indexed part of URL.
<i>totalcount</i>	Total count of words in the indexed part of URL.
<i>content_type</i>	Content-Type HTTP header returned by server.
<i>charset</i>	Document charset taken from Content-Type HTTP header or META.
<i>title</i>	First 128 characters from pages title.
<i>txt</i>	First 255 characters from page body, stripped from HTML tags.
<i>docsize</i>	Total size of URL.
<i>keywords</i>	First 255 characters from page keywords.
<i>description</i>	First 100 characters from page description.
<i>lang</i>	Not used now.
<i>words</i>	Zipped content of URL.
<i>hrefs</i>	Sorted array of outgoing href IDs from this URL.

robots

This table contains information parsed from robots.txt file for each site.

Field	Description
<i>hostinfo</i>	Host name.
<i>path</i>	Path to exclude from indexing.

sites

This table contains IDs for all indexed sites.

Field	Description
<i>site_id</i>	ID of site.
<i>site</i>	Site name with protocol, like <code>http://www.my.com/</code> .

stat

This table contains information about query statistics for each completed query.

Field	Description
<i>addr</i>	IP address of computer, from which query was requested.
<i>proxy</i>	IP address of proxy server, through which query was requested.
<i>query</i>	Query string.
<i>ul</i>	URL limit used to restrict the query.
<i>sp</i>	Web spaces used to restrict the query.
<i>site</i>	Site ID used to restrict the query.
<i>np</i>	Results page number requested.
<i>ps</i>	Results per page.
<i>sites</i>	Number of found sites matching query.
<i>urls</i>	Number of found URLs matching query.
<i>start</i>	Query processing start in seconds from UNIX epoch.
<i>finish</i>	Query processing finish in seconds from UNIX epoch.
<i>referer</i>	URL of web page from which query was requested.

subsets

Table describing all subsets, which can be used to restrict the search. Populated manually with URL masks. Subset is the set of URLs from the particular directory of site. Putting masks describing whole site is not necessary.

Field	Description
<i>subset_id</i>	ID of subset.
<i>mask</i>	URL mask. Example: <i>http://www.my.com/dir/%</i> . Examples of wrong use: <i>http://www.aspstreet.com/%</i> , <i>http://www.aspstreet/%</i> .

spaces

Table describing web spaces. Web space is the set of sites. Each site belonging to particular space must be put into separate record. Populated manually or using **-A** option of **index**. If populated manually, run **index -B** after changing this table.

Field	Description
<i>space_id</i>	ID of web space.
<i>site_id</i>	ID of site belonging to the space, refers to sites.site_id.

tmpurl

Table describing URLs indexed since start of last indexing. Used for debugging.

Field	Description
<i>url_id</i>	URL ID.
<i>thread</i>	Ordinal thread number, which indexed URL.

wordsite

Auxiliary table used when search is restricted to site pattern. Built at the end of indexing from **sites** table.

Field	Description
<i>word</i>	Word used in site name between dots.
<i>sites</i>	Array of site IDs, where this word is encountered.

citation

This table contains reverse index of hyperlinks. It is used only if **IncrementalCitations** is set to *no*.

Field	Description
<i>url_id</i>	URL ID.
<i>referrers</i>	Array of URL IDs, which have hyperlink to this URL.

BLOBS**wordurl.urls, wordurl1.urls**

Sites information, ordered by site_id.		
Offset	Length	Description
0	4	Offset of URL info for 1st site.
4	4	ID of 1st site where word is encountered.
8	4	Offset of URL info for 2nd site.
12	4	ID of 2nd site where word is encountered.
...		
$(N-1)*8$	4	Offset of URL info for Nth site, where N is the total number of sites in which word is encountered.
$(N-1)*8+4$	4	Offset of URL info for Nth site.
$(N-1)*8+8$	4	Offset of URL info end for Nth site. Must point to the end of blob or file.
URLs information. Follows sites information immediately. Offsets are counted from 0.		
Offset	Length	Description
0	4	URL ID of first site in sites info section.
4	2	Word count in this URL.
6	2	First position.
8	2	Second position.
...		
$6+(N-1)*2$	2	Nth position, where N is the total word count in the URL.
<i>Repeated with info for URLs from the same site, with ID greater than previous.</i>		
...		
<i>Repeated with info for URLs for next sites from sites info section.</i>		

urlwordsNN.words

This field contains gzipped content of URL.

Offset	Length	Description
0	4	Size of URL content before zipping or 0xFFFFFFFF if content is not zipped.
4	Zipped size	Zipped or original URL content.

wordsite.sites

This field contains array of sites/positions for word. Sorted by site IDs.

Structure of array element:

Bits	Description
24-31	Bitmap of positions, highest bit is set to 1 is word is first-level domain.
0-23	Site ID.

FILES

`/usr/local/aspseek/etc/DBType/tables.sql`

SEE ALSO

`aspseek(7)`, `index(1)`, `searchd(1)`.

AUTHORS

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

Man page by Kir Kolyshkin <kir@asplinux.ru> and Alexander F. Avdonkin <al@asplinux.ru>.