php全文检索:对MYSQL进行全文检索的PHP类库

真是好东西,但我还没研究出来如果要看这个详细介绍说明和演示请到这个地方看真很爽:
http://steven.haryan.to/php/KwIndex.html

注意只能在linux,unix下用

```
<?php

$debug = 0;

($debug) require \"Dumper.lib\"; //这个全文检索需要库文件你有吗？

function _debug {

global $debug;

$args = func_get_args;
```

```
 (!$debug) ;

echo \"<pre>debug: \", htmlentities(join(\"\", $args)), \"</pre><br>\\n\";

}


 KwIndex {


# CONSTRUCTOR

##############


function KwIndex($args) {

# check for argument type

 (!is_.gif' />(&$args))

die(\"KwIndex: constructor: syntax: KwIndex(.gif' /> \\$args)\");


# check for unknown arguments

$known_arguments = .gif' />_flip(.gif' />(

\"linkid\", \"db_name\", \"hostname\", \"username\", \"password\",

\"index_name\", \"wordlist_cardinality\", \"doclist_cardinality\",
```

```
\"stoplist_cardinality\", \"vectorlist_cardinality\",

\"max_word_length\", \"use_persistent_connection\"));

while(list($k,$v) = each($args))

 (!is($known_arguments[$k]))

die(\"KwIndex: constructor: unknown argument `$k\'\");


# required for required arguments

 (!is($args[\"db_name\"]))

die(\"KwIndex: constructor: You must specy \'db_name\'\");

 (!is($args[\"linkid\"]) &&

(!is($args[\"hostname\"]) || !is($args[\"username\"]) ||

!is($args[\"password\"])))

die(\"KwIndex: constructor: You must either specy \'linkid\' or \".

\"arguments to mysql_connect (\'hostname\', \'username\', and \".

\"\'password\')\");


# supply default values for optional arguments

 (!is($args[\"index_name\"]))
```

```
$args[\"index_name\"] = \"kwindex\";

 (!is($args[\"wordlist_cardinality\"]))

$args[\"wordlist_cardinality\"] = 100000;

 (!is($args[\"stoplist_cardinality\"]))

$args[\"stoplist_cardinality\"] = 10000;

 (!is($args[\"vectorlist_cardinality\"]))

$args[\"vectorlist_cardinality\"] = 100000000;

 (!is($args[\"doclist_cardinality\"]))

$args[\"doclist_cardinality\"] = 1000000;

 (!is($args[\"max_word_length\"]))

$args[\"max_word_length\"] = 32;

 (!is($args[\"use_persistent_connection\"]))

$args[\"use_persistent_connection\"] = 1;


#  object attributes

$this->db_name = $args[\"db_name\"];

$this->index_name = $args[\"index_name\"];

$this->wordlist_cardinality = $args[\"wordlist_cardinality\"];
```

```
$this->stoplist_cardinality = $args[\"stoplist_cardinality\"];

$this->vectorlist_cardinality = $args[\"vectorlist_cardinality\"];

$this->doclist_cardinality = $args[\"doclist_cardinality\"];

$this->max_word_length = $args[\"max_word_length\"];


 (!is($args[\"linkid\"])) {

 ($args[\"use_persistent_connection\"]) {

$linkid = mysql_pconnect($args[\"hostname\"], $args[\"username\"],

$args[\"password\"]);

} {

$linkid = mysql_connect($args[\"hostname\"], $args[\"username\"],

$args[\"password\"]);

}

 (!$linkid)

die(\"KwIndex: constructor: Can\'t connect to database: \".

mysql_error);

} {

$linkid = $args[\"linkid\"];
```

```
}

$this->linkid = $linkid;
$idx = $this->index_name;

 (!mysql_select_db($this->db_name, $linkid))
die(\"KwIndex: constructor: Can\'t select DB: \".
mysql_error($linkid));

 (!$this->_index_exists) $this->_create_index;

# load stoplist as keys of .gif' />
$this->stoplist = .gif' />;
 (!($res = mysql_query(\"SELECT word FROM ${idx}_stoplist\",
$linkid)))
die(\"KwIndex: constructor: Can\'t load stoplist: \".
mysql_error($linkid));
while($row = mysql_fetch_row($res)) {
```

```
$this->stoplist[ strtolower($row[0]) ] = 1;

}

} // constructor



# PUBLIC METHODS

#################


function &document_sub($doc_ids) {

die(\"KwIndex: document_sub: this method must be overriden\");

}


function add_document($doc_ids) {

 (!is_.gif' />(&$doc_ids))

die(\"KwIndex: syntax: add_document(.gif' /> \\$doc_ids)\");


 (!(&$doc_ids))  1;
```

```
$wordlist = .gif' />;


# structure: ( \'word1\' => [ [doc_id,freq], ... ], ... )

$doclist = .gif' />;

# format: ( doc_id => n, ... ); # n = number of words in document


# retrieve documents

####################

$docs = $this->document_sub(&$doc_ids);


(!is_.gif' />(&$docs))

die(\"KwIndex: add_document: \".

\"\'document_sub\' does not  an .gif' />\");

 ((&$doc_ids) < (&$docs))

die(\"KwIndex: add_document: \".

\"\'document_sub\' does not  enough documents\");

 ((&$doc_ids) > (&$docs))

die(\"KwIndex: add_document: \".
```

```
\"\'document_sub\' s too many documents\");


# split documents o words

############################

while(list($id, $doc) = each($docs)) {

 (!is($doc) || !strlen($doc)) continue;


$words = $this->_split_to_words($doc);

$num_of_words = (&$words);

# note: this means that numbers, etc are counted

$doclist[$id] = $num_of_words;


# filter non-qualying words: 1-char length, numbers, words

# that are too long


$w2 = .gif' />;

while(list($k, $v) = each($words)) {

$len = strlen($v);
```

```
$lower_v = strtolower($v);

 ($len > 1 &&

$len <= $this->max_word_length &&

preg_match(\"/[a-z]/\", $lower_v) &&

!is($this->stoplist[$lower_v])) $w2[ $lower_v ];

}


while(list($k, $v) = each($w2)) {

$lower_k = strtolower($k);

 (!is($wordlist[$lower_k]))

$wordlist[$lower_k] = .gif' />;

.gif' />_push($wordlist[$lower_k], .gif' />($id, $v/$num_of_words));

}

}


#_debug(\"wordlist: \", Dumper($wordlist));


# submit to database
```

```
####################

$linkid = $this->linkid;

$idx = $this->index_name;


# lock the tables in  some other process remove a certain word

# between step 0 and 1 and 2 and 3

(!mysql_query(\"LOCK TABLES ${idx}_doclist WRITE, \".

\"${idx}_vectorlist WRITE, \".

\"${idx}_wordlist WRITE\",

$linkid)) {

$this->ERROR = \"Can\'t lock tables when adding documents: \".

mysql_error($linkid);

;

}


# 0

# add the docs first

#_debug( \"doclist = \", Dumper($doclist));
```

```
while(list($k,$v) = each($doclist)) {

(!mysql_query(\"REPLACE INTO ${idx}_doclist (id,n) VALUES (\".

\"\'\".(addslashes($k)).\"\'\".

\",\".

\"\'\".(addslashes($v)).\"\'\".

\")\",

$linkid)) {

$this->ERROR = \"Can\'t add doc id=`$_\' to doclist: \".

mysql_error($linkid);

mysql_query(\"UNLOCK TABLES\", $linkid);

;

}

}



# 1

# and then add the words

while(list($k,$v) = each($wordlist)) {

(!mysql_query(\"INSERT IGNORE INTO ${idx}_wordlist (word) \".
```

\"VALUES (\".

\"\'\".(addslashes($k)).\"\'\".

\")\",

$linkid)) {

$this->ERROR = \"Can\'t add word `$k\' to wordlist: \".

2009-2-12 5:09:19
疯狂代码 http://CrazyCoder.cn/