

PESQUISA COM ACENTOS DENTRO DO MYSQL

Diego M. Rodrigues (diego@drsolutions.com.br)

O objetivo desse material é mostrar aos desenvolvedores da comunidade PHP como realizar buscas dentro de tabelas MySQL quando o conteúdo dos campos contém textos acentuados.

Um exemplo típico seria dentro de uma loja virtual de livros, buscar livros baseado no título dos mesmos.

Vamos imaginar que nosso visitante queira buscar o livro “PHP: A Bíblia”. Temos dois problemas com o acento “í” aqui. O primeiro é como isso está cadastrado no banco de dados, pois podemos ter “PHP: A Bíblia”, “PHP: A Biblia” e ainda “PHP: A Bíbliã” (caso tenha sido passado um `html_entities()` na rotina de inclusão).

O segundo problema está no formulário que o visitante vai preencher para realizar a busca, pois ele pode escrever “bíblia”, “biblia” , “BLÍBLIA”, “biblia” e por aí vai... se passarmos um `html_entities()` no conteúdo do que foi informado pelo usuário, essas coisas ainda poderiam virar algo como “bíbliã”...

A forma mais básica de realizarmos buscas de “pedaços” de campos dentro de uma tabela é usando a cláusula `LIKE` do SQL. Por exemplo:

```
SELECT * FROM livros WHERE titulo LIKE "%bíblia%"
```

Essa query irá retornar os registros da tabela livros que dentro do título exista a palavra “bíblia”... mas note que irá busca “bíblia”, e não “biblia” e nem “bíbliã”, portanto se nosso visitante buscasse por “biblia” não encontraria o livro “PHP: A Bíblia”.

A cláusula `LIKE`, embora muito útil em diversos casos, deixa muito a desejar quanto estamos lidando em buscas dentro de textos na língua portuguesa. Uma boa alternativa é o uso de Expressões Regulares usando a cláusula `REGEXP`.

A primeira pergunta que surge é: “O que são expressões regulares?”

“Bem resumido, uma expressão regular é um método formal de se especificar um padrão de texto.” (Aurélio Marinho Jargas)

A segunda seria: “Você pode me dar um exemplo?”

Poderíamos usar uma ER(expressão regular) para realizar buscas do tipo:

Todos os títulos que comecem com “PHP”

Todos os títulos que terminem com “invadir”

Todos os títulos que contenham “b” seguido de “í” ou “ï” ou “í” seguido de “bliã”

Esse artigo não visa explicar as ERs em si, mas como usá-las. Irei apenas introduzir alguns conceitos básicos a seguir e disponibilizarei links no final do artigo para os mais interessados.

Vamos lá, um pouco de ERs:

“\$” quer dizer começo de uma linha

“^” quer dizer final de uma linha

“.” quer dizer qualquer coisa

“(alb)” quer dizer a OU b

“(1|2|3)” quer dizer 1 OU 2 OU 3

“*” quer dizer tanto faz

Com isso já posso dar alguns exemplos:

“\$php” linha que comece com “php”

“\$p.p” linha que comece com “p” seguido “um caractere qualquer” seguido de “p”, poderia ser php, p2p, pgp...

“p(h|g)p^” linha que termina com “php” ou “pgp”

As ERs são muito mais do que essas poucas regras que eu passei. Os leitores mais atentos já devem estar percebendo como elas podem ser poderosas... Mais alguns exemplos:

Pegar uma data dentro de uma página

Validar um e-mail

Pegar a cotação do dólar

Voltando ao tema do artigo, vamos ver como usar uma ER dentro do MySQL. Vamos realizar a busca: todos os livros que o título comece com “php”:

```
SELECT * FROM livros WHERE titulo REGEXP "$php"
```

Outra busca seria, todos os livros que contenham a palavra php ou pgp:

```
SELECT * FROM livros WHERE titulo REGEXP "p(h|g)p"
```

Voltando ao nosso primeiro exemplo, vamos buscar “bíblia” ou “biblia” ou “bíbblia”... na verdade vamos buscar uma palavra que comece com “b”, tenha “i” ou “í” ou “iacute;”, e depois “blia”:

```
SELECT * FROM livros WHERE titulo REGEXP "b(i|í|&iacute;)blia"
```

E se estivermos falando da palavra “ação”, o usuário pode digitar “acao”, “ação”, “ação”... vamos então montar uma ER para o “ç” e para o “ã”:

```
SELECT * FROM livros WHERE titulo REGEXP "a(c|ç|&ccedil;)(a|ã|&atilde;)o"
```

Aí vem a terceira pergunta: “Como adivinhar onde estão os acentos para montar a ER?”

Não dá meu amigo, o jeito é montar uma ER bem abrangente, pois é melhor o visitante receber uma lista com alguns livros a mais do que a menos...

A minha proposta é a seguinte: Pegue o que o usuário digitou e remova os acentos, depois busque todas as possibilidades com acentos e sem acentos... por exemplo, se o usuário digitar “acao”, “ação” ou “ação”, vamos transformar em “acao”.

Agora vamos pensar.. “a” pode ser “a”, “ã”, “á”, “à”... “c” pode ser “c” ou “ç”... “o” pode ser “o”, “õ”, “ó”, “ó”... Montando a ER:

```
SELECT * FROM livros WHERE titulo REGEXP "(a|ã|á|à)(c|ç)(a|ã|á|à)(o|õ|ó|ò)"
```

Vocês já devem ter sentido como uma expressão dessas pode ficar grande. Eu uso o seguinte apenas para a letra “a”:

```
(a|ã|á|à|â|â|&atilde;|&agrave;|&agrave;|&uml;|&acirc;|Ã|Á|À|Ä|Â|&Atilde;|&Aacute;|&Aggrave;|&Auml;|&Acirc;)
```

A quarta e última pergunta seria: “Você não tem aí uma funçãozinha pronta para montar essas ERs?”

Sim tenho. Ela funciona da seguinte maneira: Primeiro converte tudo para minúsculo, depois tira todos os acentos e por último monta a ER.

```
function stringParaBusca($str) {
    //Transformando tudo em minúsculas
    $str = trim(strtolower($str));

    //Tirando espaços extras da string... "tarcila almeida" ou "tarcila almeida" viram "tarcila almeida"
    while ( strpos($str, " ") )
        $str = str_replace(" ", "", $str);

    //Agora, vamos trocar os caracteres perigosos "ã,á..." por coisas limpas "a"
    $caracteresPerigosos =
    array("ã","á","à","â","ä","é","ê","ë","è","í","î","ï","ï","ó","ô","õ","ö","ú","ù","ü","û","ç");
    $caracteresLimpos =
    array("a","a","a","a","a","e","e","e","e","i","i","i","i","o","o","o","o","o","o","u","u","u","u","u");
    $str = str_replace($caracteresPerigosos,$caracteresLimpos,$str);

    //Agora que não temos mais nenhum acento em nossa string, e estamos com ela toda em "lower",
    //vamos montar a expressão regular para o MySQL
    $caractresSimples = array("a","e","i","o","u","c");
    $caractresEnvelopados = array("{[a]}", "{[e]}", "{[i]}", "{[o]}", "{[u]}", "{[c]}");
    $str = str_replace($caractresSimples,$caractresEnvelopados,$str);
    $caracteresParaRegExp = array(
        "(a|ã|á|à|â|&atilde;|&aacute;|&agrave;|&auml;|&acirc;|Ã|Á|À|Â|Â|&Atilde;|&Aacute;|&Agrave;|&Auml;|&Acirc;)",
        "(e|é|ê|ë|ê|&eacute;|&egrave;|&euuml;|&ecirc;|É|Ê|Ë|É|&Eacute;|&Egrave;|&Euuml;|&Ecirc;)",
        "(i|î|ï|î|&iacute;|&igrave;|&iuml;|&icirc;|Í|Î|Ï|Î|&Iacute;|&Igrave;|&Iuml;|&Icirc;)",
        "(o|õ|ó|ò|ô|&otilde;|&oacute;|&ograve;|&ouuml;|&ocirc;|Õ|Ó|Ò|Ô|Ô|&Otilde;|&Oacute;|&Ograve;|&Ouml;|&Ocirc;)",
        "(u|ú|û|ü|&uacute;|&ugrave;|&uuuml;|&ucirc;|Ú|Û|Ü|Ú|&Uacute;|&Ugrave;|&Uuml;|&Ucirc;)",
        "(ç|Ç|&ccedil;|&Ccedil;)" );
    $str = str_replace($caractresEnvelopados,$caracteresParaRegExp,$str);

    //Trocando espaços por .*
    $str = str_replace(" ", ".*", $str);

    //Retornando a String finalizada!
    return $str;
}
```

Poderíamos então usar essa função e montar uma query:

```
$sql = "SELECT * FROM livros WHERE titulo REGEXP \"\" . stringParaBusca("acao") . \"\"";
$result = mysql_query($sql);
```

Quem quiser um exemplo completo, pode pegar o arquivo ZIP:

<http://www.drsoolutions.com.br/exemplos/exemplegex.zip>

Aos leitores mais interessados em expressões regulares, sugiro o site:

<http://guia-er.sourceforge.net/guia-er.html>

Aos leitores que desejam mais informações sobre como as ERs são usadas no MySQL, sugiro o manual do MySQL:

<http://dev.mysql.com/doc/refman/4.1/pt/regexp.html>

Abraços a todos,

Diego M. Rodrigues

<http://www.drsoolutions.com.br>

diego@drsoolutions.com.br

diego.rodrigues@poli.usp.br