Pesquisando por Similaridade com as funções SOUNDEX e DIFFERENCE do SQL Server

Autor : Marcelo Ramos

Data : 29/05/2009

Imagine a seguinte situação :

Você manda periodicamente um informativo por email para todos os contatos cadastrados em uma tabela de contatos, com nome e email. Só que você tem notado que muitos emails voltam por erro de digitação na hora do cadastro, produzindo algo como HOITMAIL ao invéz de HOTMAIL por exemplo. Você precisa corrigir isso, mas como fazer ?

Há duas funções no SQL que podem nos ajudar muito nesse caso. SOUNDEX e DIFFERENCE. Ambos avaliam semelhança de strings, mas vamos entender melhor na prática.

Vamos criar a Tabela de Contatos

```
CREATE TABLE DBO.CONTATOS
(
      ID          INT IDENTITY(1,1) PRIMARY KEY NOT NULL,
      NOME  VARCHAR(100) NOT NULL,
      EMAIL VARCHAR(100) NOT NULL
)
GO
```

Inserir os Dados Simulando o Problema

```
INSERT INTO DBO.CONTATOS( NOME, EMAIL )

-- HOTMAIL SEM ERRO DE DIGITAÇÃO
SELECT 'MARCELO'  , 'MARCELO@HOTMAIL.COM'       UNION ALL
-- AQUI INSERIMOS 4 VARIAÇÕES COM ERRO DE CONTAS HOTMAIL,
-- HOTMAL, HOITMAIL, HOITMAI e HOTMAIL.COM.BR
SELECT 'JOAO'     , 'JOAO@HOTMAL.COM'           UNION ALL
SELECT 'MARIA'    , 'MARIA@HOITMAIL.COM'        UNION ALL
SELECT 'JOSE'     , 'JOSE@HOITMAI.COM'          UNION ALL
SELECT 'ANTONIO'  , 'ANTONIO@HOTMAIL.COM.BR'    UNION ALL
-- AQUI MAIS 3 CONTAS DE OUTROS PROVEDORES PRA SIMULAR VARIEDADE DE
DOMINIOS
SELECT 'FULANO'   , 'FULANO@GMAIL.COM'          UNION ALL
SELECT 'BELTRANO' , 'BELTRANO@TERRA.COM.BR'     UNION ALL
SELECT 'CICLANO'  , 'CICLANO@YAHOO.COM'
GO
```

Depois disso, teremos nossa tabela populada como na imagem abaixo :

	ID	NOME	EMAIL
1	1	MARCELO	MARCELO@HOTMAIL.COM
2	2	JOAO	JOAO@HOTMAL.COM
3	3	MARIA	MARIA@HOITMAIL.COM
4	4	JOSE	JOSE@HOITMAI.COM
5	5	ANTONIO	ANTONIO@HOTMAIL.COM.BR
6	6	FULANO	FULANO@GMAIL.COM
7	7	BELTRANO	BELTRANO@TERRA.COM.BR
8	8	CICLANO	CICLANO@YAHOO.COM

Com todo o ambiente montado, agora podemos fazer nossas buscas usando SOUNDEX e DIFFERENCE. Lembrando que nosso objetivo é trazer os emails HOTMAIL e as variações

causadas por erros de digitação como HOITMAIL, HOTMAL e HOTMAIL.COM.BR para que possamos corrigir a base e diminuir os emails retornados.

Usando SOUNDEX

A sintaxe é essa :

```
SELECT * FROM TABELA
WHERE SOUNDEX( COLUNA ) = SOUNDEX( 'VALOR A COMPARAR' )
```

Para o nosso caso fica assim :

```
SELECT * FROM DBO.CONTATOS
WHERE
      SOUNDEX(
      RIGHT( EMAIL, LEN( EMAIL ) - CHARINDEX('@', EMAIL ) )
      ) = SOUNDEX( 'HOTMAIL.COM' )
```

No Script acima, `RIGHT( EMAIL, LEN( EMAIL ) - CHARINDEX('@', EMAIL ) )` nos retorna a parte do email após o @, e comparamos isso a HOTMAIL.COM

O resultado será o seguinte :

	ID	NOME	EMAIL
1	1	MARCELO	MARCELO@HOTMAIL.COM
2	2	JOAO	JOAO@HOTMAL.COM
3	3	MARIA	MARIA@HOITMAIL.COM
4	5	ANTONIO	ANTONIO@HOTMAIL.COM.BR

Repare que conseguimos o email correto e as variações com erros de digitação. Com isso podemos efetuar a correção com um UPDATE simples, fazendo uso da função REPLACE.

```
UPDATE DBO.CONTATOS
SET EMAIL = REPLACE(
                  EMAIL,
                  RIGHT( EMAIL, LEN( EMAIL ) - CHARINDEX('@', EMAIL )
),
                  'HOTMAIL.COM'
                  )
FROM DBO.CONTATOS
WHERE
      SOUNDEX(
      RIGHT( EMAIL, LEN( EMAIL ) - CHARINDEX('@', EMAIL ) )
      ) = SOUNDEX( 'HOTMAIL.COM' )
```

Após rodar nosso UPDATE, teremos os 4 registros acertados

	ID	NOME	EMAIL
1	1	MARCELO	MARCELO@HOTMAIL.COM
2	2	JOAO	JOAO@HOTMAIL.COM
3	3	MARIA	MARIA@HOTMAIL.COM
4	5	ANTONIO	ANTONIO@HOTMAIL.COM

Vamos ao mesmo exemplo usando a função DIFFERENCE.

A sintaxe :

```
SELECT * FROM TABELA
WHERE DIFFERENCE( COLUNA, 'VALOR A COMPARAR' ) = NIVEL DE SIMILARIDADE
```

O nível de similaridade é um inteiro e o valor máximo é 4 e o mínimo é 0.
No nosso exemplo :

```
SELECT *
      FROM  DBO.CONTATOS
      WHERE DIFFERENCE(
            RIGHT( EMAIL, LEN( EMAIL ) - CHARINDEX( '@', EMAIL ) )
                  , 'HOTMAIL.COM' ) = 4
```

Obtemos o mesmo resultado que usando SOUNDEX.

	ID	NOME	EMAIL
1	1	MARCELO	MARCELO@HOTMAIL.COM
2	2	JOAO	JOAO@HOTMAL.COM
3	3	MARIA	MARIA@HOITMAIL.COM
4	5	ANTONIO	ANTONIO@HOTMAIL.COM.BR

Podendo também fazer um UPDATE com REPLACE pra corrigir os erros de digitação.

```
UPDATE DBO.CONTATOS
SET EMAIL = REPLACE(
            EMAIL,
            RIGHT( EMAIL, LEN( EMAIL ) - CHARINDEX('@', EMAIL ) ),
            'HOTMAIL.COM'
            )
FROM  DBO.CONTATOS
WHERE DIFFERENCE(
            RIGHT( EMAIL, LEN( EMAIL ) - CHARINDEX( '@', EMAIL ) )
            , 'HOTMAIL.COM' ) = 4
```

Resultados corrigidos :

	ID	NOME	EMAIL
1	1	MARCELO	MARCELO@HOTMAIL.COM
2	2	JOAO	JOAO@HOTMAIL.COM
3	3	MARIA	MARIA@HOTMAIL.COM
4	5	ANTONIO	ANTONIO@HOTMAIL.COM

Lembrando que o 4 é o nível máximo de similaridade, podendo ser alterado pra que a busca fique mais flexível.

Espero que seja útil.
Aguardem os próximos artigos de SQL.

[]s
Marcelo Ramos
Arquiteto de Sistemas
MCP .NET
http://www.marceloramos.net/Blog
mramos.oliveira@gmail.com

REPLACE (Transact-SQL)
http://msdn.microsoft.com/en-us/library/ms186862.aspx

RIGHT (Transact-SQL)
http://msdn.microsoft.com/en-us/library/ms177532.aspx

SOUNDEX (Transact-SQL)
http://msdn.microsoft.com/en-us/library/ms187384.aspx

DIFFERENCE (Transact-SQL)
http://msdn.microsoft.com/en-us/library/ms188753.aspx