BUSCA FONÉTICA EM PORTUGUÊS DO BRASIL

FRED JORGE TAVARES DE LUCENA

Fred.lucena@unibratec.com.br

RESUMO

A necessidade de melhoria no acesso às informações armazenadas em grandes bases de dados tem levado as empresas a buscarem formas de tornar as consultas mais eficientes. As línguas, sejam faladas ou escritas, geram dificuldades na sua grafia quer pela ortografia e gramática oficial ou pelas influências recebidas. A língua portuguesa utilizada no Brasil também possui as suas regras ortográficas e gramaticais além de uma grande quantidade de influências internas e externas que dificultam o seu aprendizado. O estrangeirismo é muito presente e termina por criar novas palavras que são incorporadas na sua grafia original ou sofrem aportuguesamento, gerando assim, novas dificuldades e às vezes criando exceções as regras existentes. A grande extensão territorial e diversidade cultural, também são responsáveis por particularidades que podem contribuir para criar dificuldades na localização e acesso as informações. A rotina SOUNDEX[1]®, desenvolvida nos Estados Unidos em 1918, tenta minimizar os erros causados pela grafia errada das palavras a partir da codificação dos fonemas. Por ter sido desenvolvida segundo as necessidades norte americanas, a rotina SOUNDEX®, incorporada ao Oracle[2]®, apresentou um comprometimento da sua eficácia quando aplicada a língua portuguesa. Com o intuito de criar uma função capaz de abordar todas as características da língua portuguesa, foi desenvolvido um estudo que primeiramente analisou as deficiências da rotina SOUNDEX®. Após o estudo, foi desenvolvida uma análise detalhada dos fonemas da língua portuguesa e também os erros de grafia freqüentemente cometidos. Finalizados os estudos, chegou-se a função em PL/SQL[3]® que recebeu o nome de BUSCABR e que, em língua portuguesa, obteve um resultado bastante eficaz.

Palavras-chave: Consultas. Fonéticas. SOUNDEX. ORACLE. PL/SQL. BUSCABR.

[1] SOUNDEX é marca registrada de **Robert Russell** e **Margaret Odell**
[2] ORACLE é marca registrada da Oracle Corporation
[3] PL/SQL é marca registrada da Oracle Corporation

ABSTRACT

Necessity of improve access to informations stored in large data bases has lead companys to search forms to become more efficient consult methods. Spoken or written languages generates spelling difficulties, as for official grammar as for regionnal influences. Brasilian portuguese language also has orthographic and grammatical rules and has, too, a lot of internal and external influences that makes trouble on learning. The foreignism is very present in brasilian portuguese language and lead to create new words that are incorporated to its original spelling or create portuguesement that generates new difficulties and exceptions to existents rules. The large cultural diversities and territorial extension also are responsible for peculiarities that can contribute to create difficulties on access and location of informations. The SOUNDEX1® routine, developed in USA on1918 year, try to minimize spelling wrongs by using phonemes codes. Because it had been developed depending on american necessities the SOUNDEX1® routine incorporated to the Oracle2®, presented effectiveness problems when applied to portuguese language. In order to create a function able to involve all portuguese language characteristics, it was developed a research that analysed existing deficiencies in SOUNDEX1® routine. After that, detailed analysis of portuguese language phonemes associated to spelling wrongs, was developed. The finalised studies concluded for a function developed on PL/SQL3®, subsequently named BUSCABR, that obtained good results when applied in portuguese language.

Key Words: Consultas. Fonéticas. SOUNDEX. ORACLE. PL/SQL. BUSCABR.

LISTA DE TABELAS

Tabela 1 - Distribuição fonética das letras segundo o código SOUNDEX. 20
Tabela 2 - Distribuição fonética das letras segundo o código BUSCABR. 21
Tabela 3 - Classificação das consoantes segundo os critérios de emissão dos fonemas. 22
Tabela 4 - Sites de busca na internet. 23

ABREVIATURAS E SIGLAS

Cadê – Site de busca na internet, criado por Gustavo Viberti.
Google – Site de busca na internet, criado por Sergey Brin (Russia) e Larry Page (EUA).
Oracle 9i – Sistema Gerenciador de Banco de Dados da Oracle Corporation.
PL/SQL – Linguagem de programação da Oracle Corporation.
Radix – Ferramenta de busca na internet desenvolvida pelo César-UFPE.
SOUNDEX – Função de busca fonética desenvolvida por Robert Russell e Margaret Odell.

1 INTRODUÇÃO

A busca pelo aumento na velocidade da obtenção de informações tem se tornado uma constante nos dias atuais. Com a crescente integração mundial através dos diversos meios de comunicação como telefonia e internet, o volume de informações que são manipuladas diariamente cresce de forma geométrica, exigindo assim, melhorias nas diversas maneiras de acesso aos dados.

Diversas empresas, principalmente as que trabalham com grande volume de atendimento ao público, tendem a gastar um tempo maior que o necessário nas consultas as bases de dados quando têm que deduzir a grafia correta do nome de pessoas ou produtos. Hoje, qualquer perda de tempo na execução de uma tarefa, termina por gerar um custo a mais que diminui a sua competitividade no mercado.

A importância das ferramentas que facilitam a busca de informações é muito evidente hoje em dia. Um dos maiores sucessos na área de informática é o GOOGLE[4]® que tem como principal atrativo, exatamente a sua grande habilidade na busca de informações na internet. Outras empresas que prestam serviços de busca na internet, como o CADÊ[5]®, RADIX[6]® e AONDE[7]®, também apostaram no desenvolvimento de algoritmos de busca que minimizassem o tempo de resposta nas consultas a grandes bases de dados. Por outro lado, o investimento por parte das empresas proprietárias dos principais Sistemas Gerenciadores de Banco de Dados (SGBD), não tiveram a mesma preocupação. A maioria delas utiliza o mesmo sistema de busca fonética, o SOUNDEX®.

Nos Estados Unidos da América, por volta de 1918, **Robert Russell** e **Margaret Odell**, desenvolveram um sistema com o objetivo de codificar as palavras de acordo com os seus fonemas e assim facilitar a busca de informações. Esse código recebeu o nome de SOUNDEX® e tem sido usado em todo o mundo e realmente contribui para a melhoria de desempenho de muitos sistemas informatizados.

No Brasil a situação não é diferente, porém existem vários complicadores, como por exemplo, a presença de uma letra representando vários fonemas como o “X” com som de “Z” em exame, “csi” em hexa ou “cha” em enxame. Existem também, os casos em que varias letras representam o mesmo fonema como “S”, “SS” e “Ç” ou ainda “C”, “Q” e “K”. Também são encontradas palavras de origem estrangeiras que utilizam letras inexistentes no alfabeto oficial brasileiro como “K”, “W” e “Y”, ou ainda, letra que não representa fonema como a letra “H”. Além de todos os casos já citados, também existe uma série de regras e normas que também contribuem para a dificuldade do aprendizado da língua. Todos os pontos abordados geram dificuldades na obtenção de informações nas grandes bases de dados instaladas.

[4] GOOGLE é marca registrada da GOOGLE Corporation.
[5] CADÊ
[6] RADIX
[7] AONDE

Além das dificuldades inerentes a nossa gramática, também é observada as dificuldades geradas pelos regionalismos, que em um país com dimensões continentais como o Brasil, são bem maiores, apesar da existência de um único idioma. É comum, em determinadas regiões, ter a pronúncia das palavras alteradas ou parte delas suprimidas como, por exemplo, os erros gerados pelo uso ou supressão do “S”.

No caso da língua portuguesa, em especial no Brasil, as dificuldades fazem com que a eficácia da rotina SOUNDEX® seja comprometida, gerando assim, a necessidade de sua revisão ou ainda o desenvolvimento de outra rotina que possa suprir as suas limitações em relação a nossa língua.

Os trabalhos executados até então, foram focados no banco de dados Oracle® que possui como linguagem o PL/SQL®. A nossa proposta está focada no desenvolvimento de uma nova função em PL/SQL®, a partir da análise da ortografia fonêmica da língua portuguesa do Brasil, que substitua o código SOUNDEX® já que este foi concebido para a língua Inglesa e que, em virtude disto, não é eficaz em nosso idioma. Um atendente que esteja em um serviço de “help-desk” está sujeito a cometer uma série de erros causados pelos mais diversos motivos. Um deles é o mau entendimento de nomes proferidos por clientes que solicitam uma informação, seja por má dicção ou por influências regionais.

Outros são decorrentes de problemas culturais que também terminam por gerar dificuldades para quem faz uma consulta. Os bancos de dados utilizados no mercado, utilizam rotinas de consultas fonéticas baseadas no código americano SOUNDEX® e consequentemente, apresentam deficiências quando utilizado em nosso idioma.

O desenvolvimento de uma função em PL/SQL®, que contemple os erros de grafia ou pronúncia, minimizará a perda de tempo e conseqüentemente, propiciará uma maior produtividade e satisfação dos clientes já que eles não terão que ser indagados sobre a correta grafia do nome a ser consultado. Uma função de busca que contemple as deficiências nas consultas, permitirá uma melhor resposta no banco ORACLE® além de propiciar condições de sua implementação em outras linguagens utilizadas pelo mercado.

2 DESENVOLVIMENTO

Nos dias atuais qualquer perda de tempo na execução de uma tarefa termina por gerar um custo maior que o necessário, comprometendo a competitividade da empresa no mercado. O desenvolvimento de uma função em PL/SQL, que contemple os erros de grafia ou pronúncia, minimizará a perda de tempo e conseqüentemente propiciará uma maior produtividade e satisfação dos clientes já que não terão que ser indagados várias vezes sobre a grafia do dado a ser consultado. Os bancos de dados utilizados no mercado, utilizam rotinas de consultas fonéticas baseadas no código americano SOUNDEX, e consequentemente, apresentam deficiências quando utilizado em nosso idioma. A criação de uma função de busca em PL/SQL que venha a suprir essas deficiências permitirá uma melhor resposta no banco ORACLE além de propiciar condições de sua implementação em outras linguagens utilizadas pelo mercado.

Com o objetivo de investigar os principais problemas na hora de efetuar uma consulta a grandes bases de dados, testar a eficácia do código SOUNDEX® em relação à língua portuguesa e propor um novo código em substituição ao SOUNDEX®, foram utilizadas como técnicas de pesquisa os seguintes procedimentos:

- **Análise crítica sobre a eficácia do código SOUNDEX®.**

  Nesta etapa foi avaliada a eficácia do código SOUNDEX® em língua portuguesa.

- **Identificação das falhas provenientes de erros nas pronúncias das palavras.**

  Nesta etapa foram avaliadas as possíveis causas de erro de entendimento em relação a pronúncia das palavras, seja por problemas de dicção, influências regionais ou estrangeirismos.

- **Identificação das falhas provenientes de erros nas grafias das palavras.**

  Aqui foi feita uma avaliação sobre os erros que são cometidos com freqüência em virtude do desconhecimento da gramática, ou ainda, em relação às influências regionais e culturais no Brasil.

- **Proposição de um novo código em substituição ao SOUNDEX®.**

  Aqui foi proposto um novo código em substituição ao SOUNDEX®, capaz de atender, com maior eficácia, palavras escritas em língua portuguesa.

- **Teste comparativo entre o SOUNDEX® e o BUSCABR.**

  Aqui foi verificada a eficácia do BUSCABR em relação ao SOUNDEX®.

Para o estudo da ortografia fonêmica, foi utilizado o livro "Novíssima Gramática da Língua Portuguesa de Domingos Paschoal Cegalla". A existência de um estudo detalhado sobre a classificação das vogais e consoantes, segundo os critérios de modo de articulação, ponto de articulação, função das cordas vocais e função das cavidades bucal e nasal, foram essenciais ao desenvolvimento dos estudos.

Para o desenvolvimento dos trabalhos, foi utilizado o Banco de Dados ORACLE 9i® que possui a rotina interna SOUNDEX® para utilização em buscas fonéticas e o aplicativo SQL Plus[8]®.

2.1 Análise crítica sobre a eficácia do código SOUNDEX®.

Durante a analise do código SOUNDEX®, foram identificadas as seguintes deficiências:

1. **Eliminar todas as letras acentuadas;**
   Este procedimento gera um erro que pode ser constatado no seguinte exemplo: CACA e CAÇA. Ao utilizar a rotina SOUNDEX® passando-lhe primeiramente a palavra CACA, foi

[8] SQL Plus é marca registrada da Oracle Corporation.

obtido o código C000. Posteriormente, quando passado como parâmetro a palavra CAÇA, obtive-se o mesmo valor, ou seja, C000. Foi registrado que mesmo tendo fonemas bem distintos, (CACA=KAKA e CAÇA=KAÇA).

2. **Reter a primeira letra da palavra;**

   O erro gerado neste procedimento é maior que o primeiro, já que as palavras iniciadas por letra de mesmo valor fonético recebem valores diferentes como, por exemplo: VALTER e WALTER.

   Utilizando a rotina SOUNDEX® foram obtidos os códigos V436 para VALTER e W436 para WALTER. Apesar de foneticamente iguais, as palavras foram codificadas pelo SOUNDEX® como se fossem diferentes o que acarretaria em um insucesso na busca.

3. **Substituir por 0 todas as vogais e mais o H, W e Y;**

   A simples eliminação das vogais sem analisar a sua influência em relação à letra que a antecede, gera uma série de problemas. Percebeu-se que as sílabas formadas pela letra C seguida das vogais A/O/U, possuem som de K enquanto que as seguidas pelas vogais E e I têm som de Ç, ou seja, CA = KA, CE=ÇE, CI=ÇI, CO=KO, CU=KU. Exemplos: COCE e COCA.

   Utilizando a rotina SOUNDEX®, é obtido o código C200 tanto para a palavra COCE quanto para a palavra COCA o que gera uma falha na busca.

4. **Atribuir números as ocorrências de letras conforme a tabela abaixo;**

   1 = 'B', 'F', 'P', 'V'

   2 = 'C', 'G', 'J', 'K', 'Q', 'S', 'X', 'Z'

   3 = 'D','T'

   4 = 'L'

   5 = 'M','N'

   6 = 'R'

   A substituição das letras por números como ocorre na rotina SOUNDEX® limita a quantidade de fonemas que serão contemplados já que a língua portuguesa possui, segundo Cegalla, trinta e três fonemas.

2.2 Identificação das falhas provenientes de erros nas pronúncias das palavras.

Durante as investigações, foram identificados os seguintes erros de Pronúncia:

- Imperfeições na emissão das vogais e dos grupos vocálicos.

Pronúncia Correta	Pronúncia Incorreta
Umbu	Imbu
Emprego	Imprego
Ifigênia	Efigênia
Intitular	Entitular
Ouro	Ôro
Umbigo	Imbigo

- Imperfeições na emissão dos fonemas consonantais.

Pronúncia Correta	Pronúncia Incorreta
Aspas	Aspras
Bater	Batê
Bobagem	Bobage
Bugiganga	Bunginganga
Cérebro	Celebro
Chiclete	Chicrete

- Imperfeições na emissão dos fonemas ocasionados pelos regionalismos.

Pronúncia Correta	Pronúncia Incorreta
Comendo	Comeno
Dizendo	Dizeno
Pedindo	Pedino
Vamos	Vamo
Virgem	Virge
Bom	Bão

- As letras R e RR pronunciadas de forma semelhante.

Pronúncia com RR	Pronúncia com R
Carreta	Reta
Carroça	Roça
Arroba	Roupa

- A supressão de letras durante a pronúncia.

Pronúncia Correta	Pronúncia Incorreta
Perspectiva	Pespectiva
Perspicácia	Pespicácia
Superstição	Supertição

2.3 Identificação das falhas provenientes de erros na grafia das palavras.

Em relação à grafia das palávras, foram identificados os seguintes erros:

- Ao utilizar o maiúsculo e minúsculo.

Escrita Correta	Escrita Incorreta
Antônio	antônio
Carlos	carlos
Maria	maria
Pedro	pedro

- No emprego do singular e plural.

Escrita Correta	Escrita Incorreta
Ramos	Ramo
Santos	Santo
Lemos	Lemo
Chaves	Chave
Campos	Campo
Sales	Sale

- Na utilização do “Y”.

Escrita Correta	Escrita Incorreta
Yara	Iara
Yuri	Iuri

- Na utilização do “H”.

Escrita Correta	Escrita Incorreta
Henrique	Enrique
Habitação	Abitação
Habilitação	Abilitação
Hábil	Ábil

- Utilização do “ÃO” e “AO”.

Escrita Correta	Escrita Incorreta
Galpão	Galpao
Alçapão	Alçapao
Campeão	Campeao

- Emprego do “Q”, “K” ou “C”.

Escrita Correta	Escrita Incorreta
Keila	Queila
Karla	Carla
Kibutz	Quibutz

- Ao empregar o “G” e o “J”.

Escrita Correta	Escrita Incorreta
Gelatina	Jelatina
Geléia	Jeléia
Gelo	Jelo
Gemido	Jemido
Gengibre	Jengibre
Geração	Jeração
Gergelim	Jerjelim

- Utilização ou não do “C” antes do “T”.

Escrita Correta	Escrita Incorreta
Conectar	Conetar
Compactar	Compatar
Prospecto	Prospeto
Introspectivo	Introspetivo
Retrospectivo	Retrospetivo
Prospectivo	Prospetivo

- Utilização do “N” e o “M”.

Escrita Correta	Escrita Incorreta
Contem	Conten
Pedem	Peden
Dizem	Dizen
Podem	Poden
Falam	Falan
Contam	Comtam

- Utilização de “W” ou “V”.

Escrita Correta	Escrita Incorreta
Waffle	Vafle
Wagner	Vagner
Walter	Valter
Wanderson	Vanderson
Watts	Vats

- Uso do “C”, “Ç”, “S” ou “SS”.

Escrita Correta	Escrita Incorreta
Cacique	Caçique
Começo	Comesso
Comissão	Comição
Concessão	Conseção
Massa	Maça
Moção	Mossão
Mordaça	Mordassa

- Na utilização do “S”, “X”, “CH” ou “Z”.

Escrita Correta	Escrita Incorreta
Ameixa	Ameicha
Brocha	Broxa
Casa	Caza
Chuchu	Xuxu
Concha	Conxa

- No uso do “L”, “O” ou “U”.

Escrita Correta	Escrita Incorreta
Alteração	Auteração
Altura	Autura
Asfalto	Asfauto
Babel	Babeu
Caiu	Cail
Cerol	Cerou
Curau	Cural

2.4 **Proposição** de um novo código em substituição ao SOUNDEX®.

1. **Transformar em maiúsculas todas as letras da palavra a ser codificada;**

   Evita a consulta do tipo “case sensitive”, ou seja, não importa se os dados estão armazenados em letras maiúsculas ou minúsculas.

2. **Eliminar todos os acentos das vogais;**

   Elimina os problemas causados por erro de acentuação.

3. **Substituir as letras conforme tabela abaixo;**

   Evita erros gerados na grafia das palavras.

DE	PARA	DE	PARA	DE	PARA
BL, BR	B	L	R	RM	SM
CA	K	N, MD	M	RJ	J
CE, CI	S	MG	G	ST, TR, TL	T
CO, CU, CK	K	MJ	J	TS	S
Ç, CH	S	PH	F	W	V
CT	T	PR	P	X	S
GE, GI	J	Q	K	ST	T
GM	M	RG	G	Y	I
GL, GR	G	RS	S	Z	S
		RT	T		

4. **Eliminar os M, R e S no final de palavras;**

   Elimina os problemas ocasionados pelo uso inadequado do plural e do infinitivo.

5. **Eliminar todas as vogais e mais o H;**

   Elimina as vogais e o H por não terem, após as etapas anteriores, maior importância fonética.

2.5 Teste comparativo entre o SOUNDEX® e o BUSCABR.

BR=BL=B		
Broco	B620	BK
Bloco	B420	BK

CA=CO=CU=CK=K e CE=CI=Ç=S		
Casa	**C200**	**KS**
Kasa	**K200**	**KS**
Cela	**C400**	**SR**
Sela	**S400**	**SR**
Circo	**C620**	**SRK**
Sirco	**S620**	**SRK**
Coroar	**C660**	**KR**
Koroar	**K660**	**KR**
Cuba	**C100**	**KB**
Kuba	**K100**	**KB**
Roça	**R000**	**RS**
Rosa	**R200**	**RS**

CH=S (Como X=S, CH=S)		
Ameixa	**A520**	**MS**
Ameicha	**A520**	**MS**

CS=S		
TORACS	**T620**	**TR**
TORAX	**T620**	**TR**

CT=T		
Compactar	**C512**	**KMPT**
Compatar	**C513**	**KMPT**

GA=GO=GU=G e GE=GI=J e GM=M e GL=GR=G		
Gana	**G500**	**GM**
Gene	**G500**	**JM**
Gibi	**G100**	**JB**
Gostar	**G236**	**GT**
Guabiru	**G160**	**GBR**
Fleuma	**F450**	**FRM**
Fleugma	**F425**	**FRM**
Hieróglifo	**H624**	**RGF**
Hierógrifo	**H626**	**RGF**
Negro	**N260**	**MG**
Nego	**N200**	**MG**

L=R		
Luminar	**L556**	**RM**
Ruminar	**R556**	**RM**

N=M		
Mudez	**M320**	**MD**
Nudez	**N320**	**MD**

MD=D		
Comendo	**C553**	**KM**
Comeno	**C550**	**KM**

MG=G e MJ=J		
Bunginganga	**B525**	**BJG**
Bugiganga	**B225**	**BJG**

PH=F e PR=P		
Philipe	**P410**	**FRP**
Felipe	**F410**	**FRP**
Estupro	**E231**	**TP**
Estrupo	**E236**	**TP**

Q=K		
Queijo	Q200	KJ
Keijo	K200	KJ

RG=G e RS=S e RT=T		
Lagarto	L263	RGT
Largato	L623	RGT
Perspectiva	P621	PSPT
Pespectiva	P212	PSPT
Lagartixa	L263	RGTS
Largatixa	L623	RGTS

RM=SM		
Mesmo	**M250**	**MSM**
Mermo	**M650**	**MSM**

RJ=J		
Virgem	**V625**	**VJ**
Vige	**V200**	**VJ**

ST=T		
Superstição	**S162**	**SPTS**
Supertição	**S163**	**SPTS**

TR=T e TL=T e TS=T		
Estupro	**E231**	**TP**
Estrupo	**E236**	**TP**
Contrato	**C536**	**KMT**
Contlato	**C534**	**KMT**
Kubitscheck	**K132**	**KBSK**
Kubixeque	**K122**	**KBSK**

W=V		
Walter	**W436**	**VT**
Valter	**V436**	**VT**

X=S		
Exceder	**E236**	**SD**
Esceder	**E236**	**SD**

Y=I		
Yara	**Y600**	**R**
Iara	**I600**	**R**

Z=S		
Casa	**C200**	**KS**
Caza	**C200**	**KS**

3 CONCLUSÃO E CONSIDERAÇÕES FINAIS

A função BUSCABR, que fora concebida levando-se em consideração todas as características inerentes à língua portuguesa, possibilitará uma consulta mais eficiente quando utilizada em relação as existentes, uma vez que aborda de forma bastante eficaz, não só os erros de ortografia, mas também aqueles provenientes de influências lingüísticas e regionais.

Apesar de o BUSCABR ter sido desenvolvido em linguagem PL/SQL, sua implementação poderá ser feita em qualquer outra linguagem.

A língua portuguesa, como toda língua viva, sofre mudanças com o passar dos tempos e esse é um dos fatores que faz com que a função BUSCABR seja tão versátil já que permite atualização ou readequação do seu código fonte por parte do DBA.

4 REFERÊNCIAS BIBLIOGRÁFICAS

CREATIVYST. Disponível no site:
<http://www.creativyst.com/Doc/Articles/SoundEx1/SoundEx1.htm>;
Acesso em: 02/02/2006.

CREATIVYST. Disponível no site:
<http://www.creativyst.com/Doc/Articles/SoundEx1/SoundEx1.htm>;
Acesso em: 03/02/2006.

ICS. Disponível no site: <http://www.ics.uci.edu/~dan/genealogy/Miller/javascrp/soundex.htm>;
Acesso em: 03/02/2006.

MEC. Disponível nosite: <http://www.mec.gov.br/>;
Acesso em: 04/02/2006.

NATIONAL ARCHIVES. Disponível no site:
The National Archives <http://www.archives.gov/genealogy/census/soundex.html>;
Acesso em: 04/02/2006.

PASCHOAL CEGALLA, Domingos. Novíssima Gramática. São Paulo-SP.
Editora Nacional – 2002.

POR TRAS DAS LETRAS. Disponível no site:
<http://www.portrasdasletras.com.br/pdtl2/sub.php?op=gramatica/docs/empregodasconsoantes>; Acesso em: 03/02/2006.

WIKIPEDIA. Disponível no site: <http://en.wikipedia.org/wiki/soundex>.
Acesso em: 02/02/2006.

URMAN, Scott. Oracle 9i - Programação PL/SQL
Editora Campus – 2005

5 ANEXOS

Tabela 1 - Distribuição fonética das letras segundo o código SOUNDEX.

Codificação SOUNDEX	
NÚMERO	LETRAS
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

A primeira letra da palavra deve ser preservada e as letras A, E, I, O, U, H, W, e Y devem ser descartadas.

O ALGORÍTIMO

1. Transforme em maiúsculas todas as letras da palavra descartando as acentuadas.
2. Retenha a primeira letra da palavra.
3. Substitua por ‘0’ (zero) todas as ocorrências das letras 'A', E', 'I', 'O', 'U', 'H', 'W', 'Y'.
4. Substitua as demais ocorrências de acordo com a tabela abaixo:
    - 1 = 'B', 'F', 'P', 'V'
    - 2 = 'C', 'G', 'J', 'K', 'Q', 'S', 'X', 'Z'
    - 3 = 'D','T'
    - 4 = 'L'
    - 5 = 'M','N'
    - 6 = 'R'
5. Remova todos os códigos repetidos.
6. Remova todos os zeros da string.
7. Formate o código no padrão LNNN preenchendo com zero os caracteres faltantes.

Tabela 2 - Distribuição fonética das letras segundo o código BUSCABR.

Distribuição fonética das letras segundo o código BUSCABR.

DE	PARA	DE	PARA	DE	PARA
BL, BR	B	L	R	RM	SM
CA	K	N, MD	M	RJ	J
CE, CI	S	MG	G	ST, TR, TL	T
CO, CU, CK	K	MJ	J	TS	S
Ç, CH	S	PH	F	W	V
CT	T	PR	P	X	S
GE, GI	J	Q	K	ST	T
GM	M	RG	G	Y	I
GL, GR	G	RS	S	Z	S
		RT	T		

O ALGORÍTIMO

1. Converter todas as letras para Maiúsculo;
2. Eliminar todos os acentos;
3. Substituir Y por I;
4. Substituir BR por B;
5. Substituir PH por F;
6. Substituir GR, MG, NG, RG por G;
7. Substituir GE, GI, RJ, MJ, NJ por J;
8. Substituir Q, CA, CO, CU, C por K;
9. Substituir LH por L;
10. Substituir N, RM, GM, MD, SM e Terminação AO por M;
11. Substituir NH por N;
12. Substituir PR por P;
13. Substituir Ç, X, TS, C, Z, RS por S;
14. Substituir LT, TR, CT, RT, ST por T;
15. Substituir W por V;
16. Eliminar as terminações S, Z, R, R, M, N, AO e L;
17. Substituir R por L;
18. Eliminar todas as vogais e o H;
19. Eliminar todas as letras em duplicidade;

Tabela 3 - Classificação das consoantes segundo a emissão dos fonemas.

FUNÇÃO DAS CAVIDADES BUCAL E NASAL		ORAIS						NASAIS
MODO DE ARTICULAÇÃO		OCLUSIVAS		CONSTRITIVAS				OCLUSIVAS
				FRICATIVAS		VIBRANTES	LATERAIS	
FUNÇÃO DAS CORDAS VOCAIS		SURDAS	SONORAS	SURDAS	SONORAS	SURDAS	SONORAS	SONORAS
PONTO DE ARTICULAÇÃO	BILABIAIS	p	b					m
	LABIODENTAIS			f	v			
	LINGÜODENTAIS	t	d					
	ALVEOLARES			s c ç	s z	r rr	l	N
	PALATAIS			x ch	g j		lh	nh
	VELARES	c q k	g guê					

Fonte: Novíssima Gramática da Língua Portuguesa – Domingos Paschoal Cegalla

Tabela 4 - Sites de busca na internet.

Google - http://www.google.com.br
Radar Uol - http://www.radaruol.com.br/
Aonde? http://www.aonde.com/
AltaVista (EUA) - http://www.altavista.com/
A Porta - http://www.porta.com.br/
Brazilis - http://www.brazilis.com.br/
Busca (Portugal) - http://www.busca.net/
Cadê - http://www.cade.com.br/
Cusco (Portugal) - http://www.cusco.pt/
Excite (EUA) - http://www.excite.com/
Global Media - http://www.globalmedia.com.br/
Guia Útil - http://www.guiautil.com/
InfoBrasil - http://www.infobrasil.com/
Jaguatirica - http://www.jaguatirica.com.br/
Lycos - http://www.lycos.com.br/
Mercosul - http://www.mercosulsearch.com.br/
MSN - http://www.msn.com.br/
Netscape - http://www.netscape.com.pt/
O Site - http://www.osite.com.br/
Aol - http://www.americaonline.com.br/
Surf - http://www.palavrarte.com/Banca_Poesia/www.surf.com.br
Yahoo Brasil - http://www.yahoo.com.br/
Star media - http://www.starmedia.com.br/
Sapo (Portugal) - http://www.sapo.pt/
Zeek - http://www.zeek.com.br/